DOCUMENT RESUME

ED 098 828 FL 006 620

AUTHOR Chafe, Wallace L.
TITLE An Approach to Verbalization and Translation by Machine. Final Report.
INSTITUTION California. Univ., Berkeley. Dept. of Linguistics.; Rome Air Development Center, Griffiss AFB, N.Y.
REPORT NO RADC-TR-74-271
PUB DATE Oct 74
NOTE 122p.

EDRS PRICE MF-$0.75 HC-$5.40 PLUS POSTAGE
DESCRIPTORS *Artificial Intelligence; Cognitive Processes; *Computational Linguistics; *Computer Programs; Concept Formation; English; *Information Processing; Japanese; Lexicology; *Machine Translation; Models; Syntax
IDENTIFIERS *Verbalization

ABSTRACT

The report documents performance on a 24-month R&D effort oriented toward the development of a computerized model for machine translation of natural languages. The model is built around a set of procedures called verbalization, intended to stimulate the processes employed by a speaker or writer in turning stored information into words. Verbalization is seen to consist of subconceptualization and lexicalization processes which involve creative choices on the part of the verbalizer, together with algorithmic syntactic processes determined by the language being used. Translation is viewed as (1) the reconstruction of the verbalization processes which went into the original source language text and (2) the application of parallel verbalization processes in the target language. The target language verbalization looks for creative choices to the source language verbalization and tries to apply corresponding choices simultaneously with application of syntactic processes dictated by the grammar of the target language. Verbalization and translation processes are illustrated in some detail with examples taken from English and Japanese. Some of these processes have been implemented in an interactive program on CDC 6600 at the Lawrence Berkeley Laboratory (AEC), but the main intent of the report is to demonstrate the kinds of processes that need to be incorporated in such a system. (Author)

RADC-TR-74-271
Final Report
October 1974

AN APPROACH TO VERBALIZATION AND TRANSLATION BY MACHINE

The University of California at Berkeley

Rome Air Development Center
Air Force Systems Command
Griffiss Air Force Base, New York

Foreword

This Final Technical Report was prepared by the University of California at Berkeley, Department of Linguistics, Berkeley, California under Contract F30602-72-C-0406, Job Order Number 45940805 for Rome Air Development Center, Griffiss Air Force Base, New York. The work performed covered the period from 1 June 1972 through 31 May 1974.

Zbigniew L. Pankowicz (IRDT) was RADC Project Engineer.

This report was written by Wallace L. Chafe, with the collaboration of other members of the Contrastive Semantics Project. Associated with the project during its entire life were Patricia M. Clancy, Leonard M. Faltz, Christopher Murano, Nasmig Seropian. Also active during more than half of this period were Masayoshi Shibatani and Linda Sobek. Associated during shorter periods of time were Teresa M. Chen, Charles J. Fillmore, Robert E. Gaskins, and Marie-Claude Jorland. Masayoshi Hirose served as a consultant on Japanese during the last two months of the project.

This report has been reviewed by the Office of Information (OI), RADC, and approved for release to the National Technical Information Service (NTIS).

This report has been reviewed and is approved.

APPROVED: Zbigniew L. Pankowicz

ZBIGNIEW L. PANKOWICZ
Project Engineer

APPROVED: Howard Davis

HOWARD DAVIS
Technical Director
Intelligence & Reconnaissance Division

FOR THE COMMANDER: James G. McGinnis

JAMES G. McGINNIS
Lt Col, USAF
Deputy Chief, Plans Office

3/

Table of Contents

EVALUATION

The report documents results of a 24-month research effort that was directed at designing a computerized model for machine translation of natural languages. The model is conceptually based on simulation of mental activities involved in verbalization (conversion of stored knowledge into linguistic patterns of a source language) and translation (reconstruction of source language verbalization and application of parallel verbalization in a traget language). The target language verbalization consists in selection of equivalents to the source language verbalization, combined with application of syntactic conventions required by the grammar of the target language.

The effort documented in this report is a direct continuation of research on semantic and post-semantic processes, carried out over the past several years by Dr. Wallace L. Chafe and described in his book, "Meaning and the Structure of Language", University of Chicago Press, 1970. The semantic component is postulated in Dr. Chafe's work as the basis of the theory of language. This position constitutes a radical departure from the modern structuralist and transformationalist trends largely concerned with the syntactic component. Since translation is traditionally defined as a transfer of meaning from the linguistic pattern of a source language into that of a target language, machine translation R&D has to account for the semantic component in order to supply the deficiencies of second generation MT models based on lexical and syntactic aspects of natural languages.

ZDZISLAW I. PANKOWICZ
Technical Evaluator

Abstract

This report describes a model for machine translation developed at Berkeley during 1972-74. The model is built around a set of procedures called verbalization, intended to simulate the processes employed by a speaker or writer in turning stored knowledge into words. Verbalization is seen to consist of subconceptualization and lexicalization processes which involve creative choices on the part of the verbalizer, together with algorithmic syntactic processes determined by the language being used. Translation is viewed as (1) the reconstruction of the verbalization processes which went into the original source language text and (2) the application of parallel verbalization processes in the target language. The target language verbalization looks for creative choices to the source language verbalization and tries to apply corresponding choices, at the same time that it applies syntactic processes dictated by the grammar of the target language. Verbalization and translation processes are illustrated in some detail, with examples taken from English and Japanese. Some of these processes have been implemented in an interactive program using the facilities of the Lawrence Berkeley Laboratory, but the main intent of the report is to demonstrate the kinds of processes that need to be incorporated in such a system.

I. Overview

Central to the view of translation that will be presented here is the notion of verbalization. Verbalization is the application of processes by which some holistic conceptual chunk, recalled from memory, is converted into sentences and words--into a phonetically or graphically communicable linguistic representation. Such a notion assumes that the underlying content of what is being communicated is not, or need not be, in verbal form to begin with. At the very least it may be a complex system of discrete elements and relations, representable perhaps as a network of nodes and arcs. It may also involve an important nondiscrete or analog component, representable only in some other terms. Whatever may turn out to be the case here, it seems clear that some sorts of processes must be applied in order to transform the original form of storage into a verbal output: that the stored material must be verbalized.

In any particular instance of translation there are two instances of verbalization. One is the original verbalization performed by the creator of the source language text. The other is the verbalization produced in the target language by the translator. Besides being in different languages, these two verbalizations are fundamentally different in one other respect. The source language verbalization is, we might say, autonomous. It is freely produced by the speaker or writer in any way he

decides is appropriate to the content and the occasion, provided he adheres to the rules of his culture and the language he is using. The target language verbalization, on the other hand, is parasitic on the source language one. Not only must the translator adhere to the rules of his own language, he must also produce a verbalization that communicates, so far as possible, the same underlying content or knowledge that was communicated by the source language verbalization. The verbalization in the target language is thus subject to this special kind of constraint. Its producer is not free to "say what he wants," but must insofar as possible say the same thing as the producer of the source language text. We suggested in an earlier report that there are two dimensions of high quality translation, which we termed naturalness and fidelity. Naturalness is achieved when the target language verbalization adheres to all the constraints of that language; the output will then sound "natural". Fidelity is achieved to the extent that the target language verbalization communicates the same content as the source language one.

Verbalization in general, as we see it, consists of a mixture of two kinds of processes: those which necessitate creative decisions on the part of the verbalizer and those which do not, being governed by the constraints imposed by the language. We might speak of creative processes and algorithmic processes. Creative processes are ultimately governed by the content which underlies the verbalization; the verbalizer has

to decide how best to verbalize that content. Normally a range of choices will be open to him, and he must decide what will most effectively convey what he has in mind. After he has made such choices, there are often automatic consequences which follow from them because of the particular rules of the language (but which are themselves likely to lead to the necessity of further creative choices). We can say, then, with respect to the two verbalizations involved in a translation, that the producer of the source language verbalization has applied both creative and algorithmic processes, whereas in the target language verbalization only algorithmic processes are autonomously applied, the necessary creative choices being determined by the choices that were made in the source language verbalization. Thus the naturalness of the final translation depends on adherence to the algorithmic processes of the target language, while its fidelity depends on the extent to which the translation has been able to incorporate creative choices that correspond to those originally applied in the source language. In all probability there are cases where exact correspondence in these choices is not possible, and where a certain amount of autonomous creativity has to be introduced into the target verbalization as well. These are the cases where automatic translation becomes most problematic. One useful goal of machine translation research ought to be to determine precisely the nature and extent of such cases.

We are led, then, to the general picture of translation

which is shown in Figure 1. The two vertical columns represent the two verbalizations which are involved: on the left the source language verbalization and on the right the target verbalization. The input to a translation procedure, of course, is an already produced verbal output or text in the source language. The first major component of the translation procedure will have to be the reconstruction from that text of the verbalization processes by which it was produced, a kind of "de-verbalization". We will refer to this as the parsing component, although it is clearly different from conventional parsing. It aims to reconstruct, not a single deep structure underlying the surface text, but rather a series of processes by which that text was created from the knowledge--not only nonverbal but possibly even nondiscrete--which the speaker or writer had in mind. The output of the parsing component is ideally a complete reconstruction of both the creative and the algorithmic processes which the source language verbalizer applied.

The other major component of the translation procedure is the translation component. It is equivalent to a verbalization in the target language. The processes which make up this verbalization are, to the extent that they are algorithmic, those which express target language constraints and, to the extent that they are creative, those which correspond to choices already made in the reconstructed source language verbalization. The necessity of reference to the source language verbalization for creative choices at many points is suggested in Figure 1 by

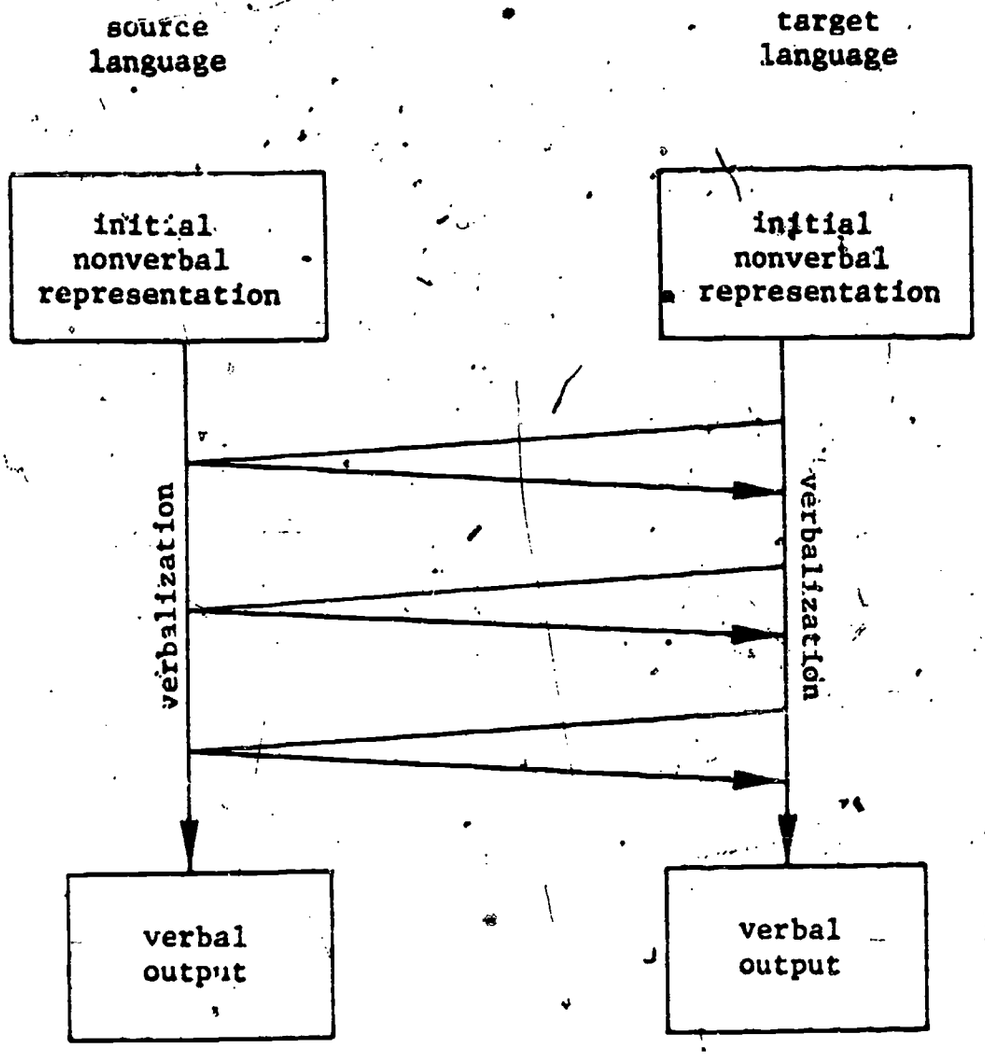
source language
target language
initial nonverbal representation
initial nonverbal representation
verbalization
verbalization
verbal output
verbal output

Figure 1

the zigzag arrows.

We believe that this picture provides a plausible basis for translation research, but needless to say it presents many problems whose solutions are only dimly foreseen at the present time. Our project has so far concentrated more of its attention on verbalization itself than on parsing or translation, since both of the latter depend on a prior understanding of verbalization. Any other ordering of priorities would be putting the cart before the horse. Any detailed investigation of the parsing component would be futile if we did not know what sort of output we would expect that component to produce; the processes that went into a particular verbalization. The translation component is a verbalization, though one ~~of a special~~ sort, and there again a detailed understanding of verbalization processes is necessary. This report, then, will be most concerned with the nature of verbalization. We will also devote considerable space to the nature of that special sort of verbalization which is translation. We will have the least to say about parsing.

For about the last nine months of the project we have been concerned with the development of an interactive computer program that will implement the verbalization processes we hypothesize. Although this program is still primitive, the intention is that it will gradually achieve increased sophistication in its ability to simulate verbalization, translation, and parsing.

As it presently simulates the processes of verbalization, it begins with an item that represents the initial holistic idea which the speaker or writer of a text wishes to communicate. It then asks the user, seated at a teletype, to make the series of creative choices that are necessary in the production of the final text. At the same time it attempts to apply on its own the algorithmic processes which are called for. It knows when creative choices are necessary, but not what choices to make. The user must decide. But it should be able to apply the algorithmic processes without help. As it simulates translation it will be able to apply the algorithmic processes of the target language automatically, and also to apply certain creative processes on its own by looking at the source language verbalization to see what creative choices were made there. Whenever it is not able to make a creative choice, the program asks the user to do so. We find that this kind of machine-user interaction provides a valuable research technique. Taking as our ultimate goal the eventual elimination of the user from the translation program altogether, we are starting with a situation in which the user intervenes at many points. As we learn more we will gradually give the machine more to do and the user less. This technique can be followed not only in verbalization and translation, but also in parsing. Whether the user will eventually disappear from the picture altogether is impossible to predict at this point.

However that may be, the goal of a program in which the

contribution of the user is significantly diminished in relation to that of the machine seems workable. Short of the final goal of eliminating the user altogether, an intermediate goal identifiable as "human-aided" machine translation can more easily be foreseen. Here the machine will do the many things for which it is suited, but a human brain will be introduced at those points where the machine has reached its limits. This intermediate goal has, we believe, significant practical as well as theoretical value.

II. Subconceptualization

We assume that a speaker or writer begins with a single, unitary, holistic conceptual chunk that he has recalled from memory and has decided, for some reason, to communicate. Thus he may have in mind some incident in which he was involved, something of interest he was previously told about or read about, some experiment he wishes to report on, or whatever. We label such a chunk, as well as the smaller chunks into which it will be analyzed, with the prefix CC (for "conceptual chunk") followed by a four-digit number. The first digit indicates the language in which verbalization is to take place ("1" for English and "2" for Japanese), and the remaining three digits constitute an arbitrary index for the particular chunk. Thus CC-1001 might be the name given to some particular chunk of this sort that is about to be verbalized in English.

We assume, furthermore, that while this chunk is from one point of view a unit, from another point of view it has a more or less rich content, and that it is this content which the speaker wishes to convey to his audience. Sometimes, though not in most cases, the initial chunk itself may have a linguistic label. If it is a folktale, for example, it may have a name like "Cinderella" or "The Three Bears". But someone who has decided to tell a story is not likely to say just "Cinderella" and let it go at that. (One is reminded of the old story

about a convention of comedians at which people said things like "49" or "178" and elicited laughter each time because everyone knew the jokes these numbers stood for.) Normally it is necessary instead for the speaker to get inside the content of this initial unit--to analyze it into smaller chunks. This kind of process can be pictured as shown in Figure 2, where the initial chunk CC-1001 has been, as we say, subconceptualized into chunks CC-1002 and CC-1003. In a text of any size each of these smaller chunks will be further broken down into still smaller ones, and so on, so that a hierarchical structure of successively smaller subconceptualizations emerges.

Subconceptualization belongs to the class of verbalization processes which are creative. Normally a chunk does not automatically determine a particular subconceptual breakdown, but the speaker must creatively choose how to subconceptualize each one. It is useful to think of the content of each chunk--each circle in Figure 2--as if it were a mountainous landscape, with the most salient aspects standing out in bold relief and the less salient appearing as only minor hills. All other things being equal, the more salient some aspect of the total content is, the more likely the speaker is to express it when he subconceptualizes. He is not likely to make exactly the same subconceptual breakdown each time he communicates the same initial chunk, partly because he may judge different things to be salient in different contexts and partly because the landscape itself may change over time, the relative salience of its dif-

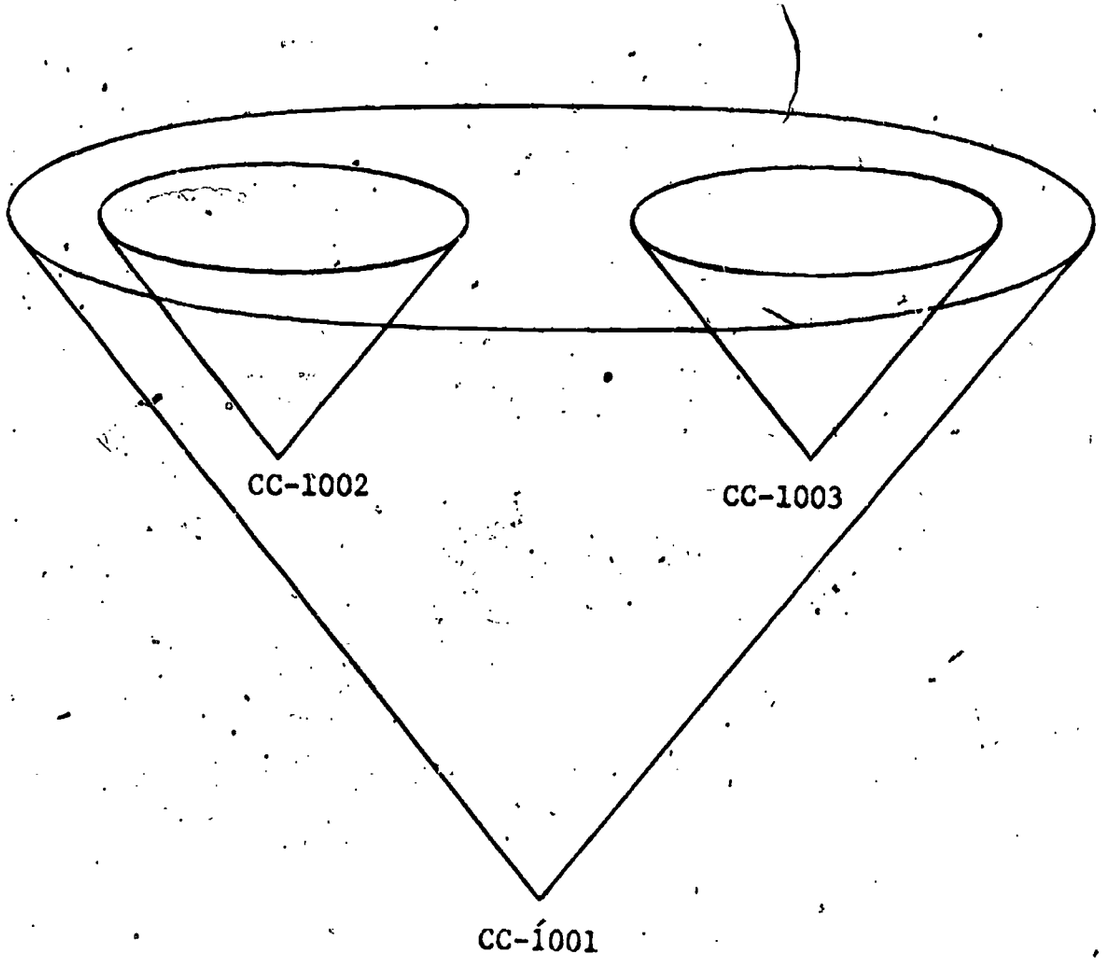

Figure 2

ferent aspects being modified in long-term memory. We assume that any particular subconceptualization necessarily leaves out part of the content of what is being subconceptualized, as suggested by the area that lies within the larger circle but outside the two smaller circles in Figure 2. Subconceptualization, that is, is necessarily a selective process. No one ever says everything he could say about what he has in mind.

Subconceptualization of a particular chunk, say CC-1001, produces two or more new chunks, say CC-1002 and CC-1003. These new chunks, furthermore, are conceived of as related to each other in some way.. For example, CC-1002 might be the "reason" for CC-1003. Suppose the entire text consisted of the sentences, "I bought a bike yesterday. I decided I need more exercise." Let us say that the first sentence is a verbalization of CC-1003 and the second sentence of CC-1002. We can say that CC-1002 is the reason for CC-1003. We write a subconceptualization process of this kind in the following way:

1) CC-1001 S> CJ-REASON (CC-1002, CC-1003)

This statement says that the initial chunk, CC-1001, is subconceptualized (S>) into the chunks CC-1002 and CC-1003, and that these two new chunks are related by the predicate labeled CJ-REASON. The prefix CJ stands for "conjunction" (derived from the grammatical, not the logical use of this term). Any relation between CC's is labeled with this prefix.

We use a different notation to represent each of the various stages in the verbalization process. At the outset, in this example, the initial chunk CC-1001 was all that was present. This initial representation, before any verbalization processes had been applied, was simply:

2) CC-1001

After the subconceptualization specified in 1) was applied, the representation became:

3) CJ-REASON
   CC-1002
   CC-1003

Subconceptualization processes are thus rewrite rules, which replace one stage in a verbalization with a subsequent stage. The format we use to represent such stages, as in 3), shows predicates with their arguments written indented below them.

In simulating verbalization our program presently asks the user to specify all the creative choices, restricting its own contribution to the application of algorithmic processes determined by the grammar of discourse, sentences, and words in the language involved. The program is labeled VAT (for "verbalizer and translator"), and we can illustrate conversations between VAT and the user identifying them as V and U respectively. The program begins by asking:

4) V: WHAT VAT TASK DO YOU WANT PERFORMED?

to which one possible answer is:

5) U: VERBALIZE CC-1001

Skipping several steps to illustrate only the rough outlines of subconceptualization, we are interested just now in the question:

6) V: HOW IS CC-1001 SUBCONCEPTUALIZED?

to which a possible answer is:

7) U: REASON (CC-1002, CC-1003)

At this point VAT will construct the representation shown in 3).

VAT will now apply an algorithmic or, as we say, syntactic process triggered by the presence of CJ-REASON in 3). The process applied is of a type that is not yet clearly understood, but we may view what we do at present at a first approximation. At the moment VAT simply takes the two CCs related by CJ-REASON and orders them so that the second will be expressed before the first. That is, for example, if CC-1002 is eventually going to be verbalized as "I decided I need more exercise" and CC-1003 as "I bought a bike yesterday", we want the two sentences to be expressed with CC-1003 preceding CC-1002. Thus VAT will automatically change the representation in 3) to the following:

8) CC-1003
CC-1002

This kind of representation, in which no predicate is shown

above the two CCs, indicates that they (or their eventual verbalizations) are to occur in the final text in the order shown, with CC-1003 preceding CC-1002.

In Japanese the corresponding syntactic process will typically lead to the attachment of CJ-"KARA" at the end of the second sentence. Thus if a representation like that in 3) were produced in a Japanese verbalization VAT would automatically change it to:

9) CC-1003
   CC-1002
   CJ-"KARA"

The quotation marks around "KARA" indicate that this is an item which will actually appear as a word in the text. Quotation marks are used for items that have a surface lexical representation. The representation in 9) is deficient in that it fails to show that CJ-"KARA" will be part of the same sentence as CC-1002, whereas CC-1003 will (or is likely to) form a different sentence. We indicate sentence boundaries with the notation CJ-".", since the period will appear in the final text. Thus fuller versions of 8) and 9) are, respectively:

10) CC-1003
    CJ-"."
    CC-1002
    CJ-"."

11) CC-1003
    CJ-"."
    CC-1002
    CJ-"KARA"

CJ-"."

The creation of these periods is a housekeeping task that need not be described in detail here.

Given a representation like that in 10), VAT will go on to ask about the subconceptualization of the first CC in the ordering. The general principle followed here is one of "depth first", in the sense that earlier items in the text are completely verbalized before the verbalization of later items is begun. This procedure probably has some psychological validity; that is, a speaker is likely to think of later parts of what he is going to say only in terms of the most general chunks, while he is elaborating the earlier parts in detail. Only after he has finished the verbalization of these earlier parts will he turn his attention to a full verbalization of the later ones.

Thus, omitting various considerations not as yet discussed, subconceptualization proceeds interactively in the following fashion:

12) V: WHAT VAT TASK DO YOU WANT PERFORMED?

U: VERBALIZE CC-1001

(VAT creates the following representation:)

CC-1001

V: HOW IS CC-1001 SUBCONCEPTUALIZED?

U: REASON (CC-1002, CC-1003)

(VAT creates first the following representation:)

CJ-REASON
CC-1002
CC-1003

(and immediately applies a stored syntactic algorithm that changes it to:)

CC-1003
CC-1002

V: HOW IS CC-1003 SUBCONCEPTUALIZED?

etc.

In this fashion a subconceptual hierarchy of any degree of complexity can be constructed and expressed.

The organization of a text may not be entirely hierarchical, however. Not only does a speaker break down larger chunks into smaller chunks--larger "concepts" into subconcepts; one chunk may also remind him of another, so that the organization which results may be in part concatenative. We have been viewing concatenation in terms of excursions away from the main hierarchy, and have been calling such excursions digressions. In some discourse, however, there is no necessary constraint that the main hierarchy be returned to, and the result may be a rambling text in which digression is added to digression. In a more tightly organized text digressions are more likely to appear as parenthetical remarks: brief sidepaths which quickly return to the main hierarchy. We use the term parenthesis for this brief and transient kind of digression.

If subconceptualization can be represented in terms of a

tree diagram (which does not, however, provide a convenient means of showing the relations between subconcepts, like CJ-REASON), then digressions can be pictured as subtrees attached to the main tree at one point or another, as suggested in Figure 3.

One other important modification of the strictly hierarchical model of subconceptualization results from the common occurrence of summarization. It is frequently the case in verbalization that an initial chunk will be subject to two separate hierarchies of subconceptualization, one of which can be identified as a summary of the other. It is characteristic of a summary that its subconceptualization processes never proceed beyond some relatively large chunks--chunks which package a relatively large content. We can contrast a subconceptualization hierarchy which is a summary with one which constitutes the body of the text and consists of subconceptualization processes which produce a larger number of chunks of smaller size.

A summary is typically expressed at the beginning or end of a text; that is, preceding or following the body. Various conventions for summaries are associated with different genres of writing. For example, a scientific article may begin with the self-conscious kind of summary that is called an abstract; a news report typically contains an opening paragraph

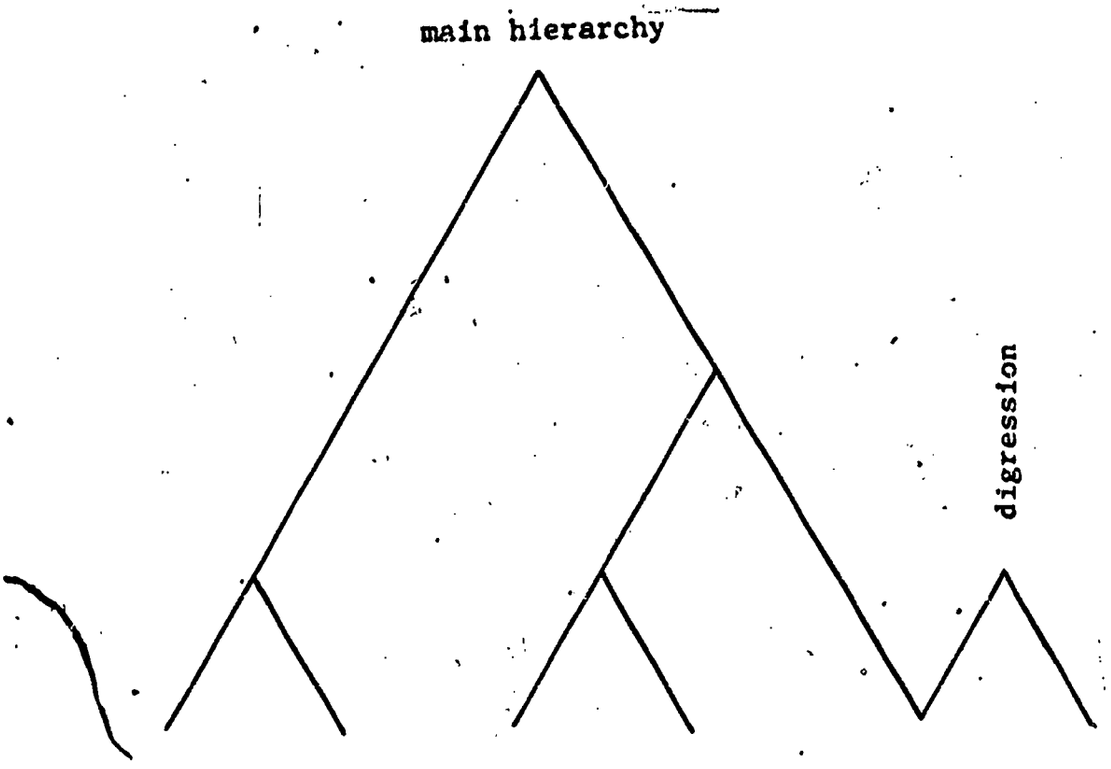

Figure 3

telling who, what, where, and when; a fable is likely to end with a moral, and so on. Our program at present simply asks, for the initial CC, whether it has an initial summary (one expressed at the beginning of the text). If the answer is yes it asks first for a subconceptualization of the summary, and moves on to ask about the body of the text only after the summary has been completely verbalized. At the end of the text it asks whether there is a final summary.

Creativity within a discourse is likely to be limited by the genre to which the discourse belongs. It would appear that there is a continuum ranging from maximally stereotyped to maximally creative discourse. Most stereotyped are those forms of discourse, such as rituals, in which the speaker has very little choice as to what he is going to say or how he is going to say it. With such discourse the "grammar" of the genre provides many of the answers to the questions VAT would otherwise have to ask the user. In other words, VAT should be able to produce ritual texts with a minimum amount of recourse to creative decisions. At the other extreme are forms of discourse such as descriptions of unique personal experiences which have never been described before, where the speaker is relatively free to make a great variety of creative decisions.

We believe it would be of considerable interest to incorporate into the verbalization process the constraints imposed by several different genres, but we have not as yet done this. As

it now stands our program does ask WHAT IS THE GENRE? as soon as it has established that a verbalization is to be performed. Possible answers that we hope to implement in the future are, for example, NEWS REPORT, PSYCHOLOGY ARTICLE, FABLE, and the like.

III. An Example

An example of these procedures as applied to a real text can be based on the following United Press report taken, slightly condensed, from the San Francisco Chronicle of May 16, 1974:

13) 1. An 11-year-old boy using a new "super-glue"
2. accidentally glued his eye shut
3. while building a model airplane,
4. and a doctor had to reopen the eye surgically.
5. Mike Harris said
6. he rubbed his left eye
7. after several drops of the glue squirted into it last Sunday
8. and found his eyelid would not move.
9. An eye surgeon debated briefly about
10. using a super glue solvent
11. but decided against it
12. for fear it might damage the boy's eye.
13. The surgeon, who asked not to be identified,
14. finally put Mike in the operating room,
15. trimmed Mike's eyelashes,
16. then opened the eyelid surgically.
17. Mike was released from the hospital Tuesday.

It is approximately the case that each of the numbered lines in

this text expresses a terminal subconcept. We assume that the text contains a number of intermediate subconcepts as well, which need to be elucidated in a subconceptual hierarchy.

Let us suppose that the combination of VAT and the user are attempting to simulate the verbalization processes that went into the production of this text. For the moment we are concerned only with subconceptualization processes (and associated syntactic algorithms). Many of the user's answers in the following conversation with VAT are intuitively based. The success of our eventual parsing component will depend on the extent to which these intuitive answers can be predicted from the text together with whatever items of background knowledge are relevant. The example will be carried only far enough to suggest the nature of the procedure.

The exchange begins in the usual way:

14) V: WHAT VAT TASK DO YOU WANT PERFORMED?

U: VERBALIZE CC-1001

VAT creates the following representation, including a text-final period:

15) CC-1001
CJ-"."

VAT's next question seeks to establish what genre constraints apply in this text:

16) V: WHAT IS THE GENRE?

U: NEWS REPORT

VAT will now assume that the text is a typical news report which begins with a summary. Its first questions will deal with the subconceptualization of the summary (expressed in the text in sentences 1-4):

17) V: HOW IS CC-1001 SUBCONCEPTUALIZED IN THE SUMMARY?

U: YIELDS (CC-1002, CC-1003)

The user has answered that the first breakdown of the summary is into two subconcepts, CC-1002 (to be expressed as lines 1-3) and CC-1003 (line 4). Furthermore, the relation between these two CCs has been identified as one labeled YIELDS, in which the first CC "leads to" or "results in" the second. YIELDS differs from another, similar relation which is labeled CAUSE in that the event conceptualized by the second CC is not a necessary consequence of the first. It is, however, something that presumably would not have happened if the event conceptualized by the first CC had not taken place. As a result of the user's answer in 17) VAT first creates the representation:

18) CJ-YIELDS
CC-1002
CC-1003
CJ-"."

and immediately applies a syntactic process which changes it to:

19) CC-1002

CJ-", AND"
CC-1003
CJ-"."

That is, the two CCs are to be expressed with the "yielder" preceding the "yielded", and they are to be connected with a comma followed by the word "AND". This is not the only way in which YIELDS can be realized, but for the sake of the example we may regard it as such. VAT will now proceed to ask about the subconceptualization of the earliest CC in 19):

20) V: HOW IS CC-1002 SUBCONCEPTUALIZED IN THE SUMMARY?
U: FRAMES (CC-1004, CC-1005)

The user has answered that CC-1002 is broken down into two CCs, CC-1004 (line 3) and CC-1005 (lines 1-2). They are related by FRAMES, a temporal relation in which the first CC occupies a time period larger than and including the time period of the second. In other words the time period of sentence 3 includes that of sentences 1-2. VAT creates, sequentially, the following two representations:

21) CJ-FRAMES
CC-1004
CC-1005
CJ-", AND"
CC-1003
CJ-"."

22) CJ-"WHILE"
CC-1004
CC-1005
CJ-", AND"
CC-1003
CJ-"."

Although there may be several possibilities for the expression of FRAMES, VAT has assumed here that two factors are involved: an ordering of the two CCs so that the "framer" precedes the "framed", and a prefixing of the word "WHILE" to the first CC. In this example the ordering of these two CCs will be reversed in a subsequent operation.

We would expect VAT to ask next about the subconceptualization of CC-1004, but by a means not yet discussed VAT will discover that this is a terminal CC (one not further subconceptualized). If CC-1004 were followed by "." or by ", AND", VAT would proceed to ask questions directed at the complete verbalization of this CC. But since CC-1004 is not followed by one of these boundaries, attention is next focused on CC-1005:

23) V: HOW IS CC-1005 SUBCONCEPTUALIZED IN THE SUMMARY?

U: FRAMES (CC-1006, CC-1007)

VAT creates the following representation:

24)
```
CJ-"WHILE"
CC-1004
CJ-FRAMES
   CC-1006
   CC-1007
CJ-", AND"
CC-1003
CJ-"."
```

The user has said that CC-1006 (line 1) occupies a time period which includes CC-1007 (line 2). So far we would expect this second instance of FRAMES to be expressed by prefixing the word

"WHILE" to CC-1006, as was done in 22). Let us suppose, however, that FRAMES actually triggers a more complex algorithm which says in effect that one "WHILE" in a sentence is enough, and that a second instance of FRAMES will lead to a different expression. Here the second instance leads to the creation of a relative clause which will modify one of the constituents of CC-1007. Furthermore, the already created "WHILE" clause will be moved to a position after CC-1007. (This ordering of the CCs does appear to be maximally natural. It would be slightly less desirable, for example, to produce "While he was building a model airplane an 11-year-old boy, using a new "super-glue", accidentally glued his eye shut." Certainly, however, the differences in this area are very subtle.). We will indicate the relative clause status of CC-1006, to be embedded within the expression of CC-1007, with a slash notation:

25) CC-1007 / CC-1006
CJ-"WHILE"
CC-1004
CJ-", AND"
CC-1003
CJ-"."

The representation in 25) will be discovered to be the final one in the subconceptualization of the summary, which has been found to consist of four CCs (ultimately four clauses) joined together in the manner indicated. VAT will now proceed to verbalize the summary completely, making use of other kinds of processes. When that has been done, it will say:

26) V: WE NOW MOVE TO THE BODY OF THE TEXT. HOW IS CC-1001 SUBCONCEPTUALIZED?

U: YIELDS (CC-1002, CC-1003)

This is, of course, the same answer that was given to the corresponding question in 17) above. As CC-1002 and CC-1003 are further elaborated, however, many differences will emerge. Ultimately CC-1002, which was expressed in sentences 1-3 of the summary, will be expressed in the body of the text in sentences 5-8. CC-1003, expressed in the summary as sentence 4, will be expressed in the body in sentences 9-17.

We will not repeat here the operations involved in the subconceptualization of the body of the text. They are for the most part similar to those illustrated above. Various other relations between CCs are introduced: for example, that between CC-1015 (lines 9-12) and CC-1016 (lines 13-16). The first of these CCs involves an alternative that is rejected in favor of the alternative conceptualized in the second; thus, the relation may be labeled REJECTED-IN-FAVOR-OF. Within CC-1015 there is a relation of CONCESSION (denial of expectation) between CC-1017 (lines 9-10) and CC-1018 (lines 11-12). It will be of considerable interest to isolate relations of this sort in a variety of texts, and to determine the ways in which they may be expressed under varying circumstances in different languages.

The text does contain one example of a parenthesis, expressed in the nonrestrictive relative clause in line 13. The

fact that the surgeon asked not to be identified is a minor digression from the mainstream of the account. It is attached to the node representing the surgeon which will become a constituent of CC-1022 (lines 14-16).

IV. Lexicalization of a CC

We use the term lexicalization to refer to another major component of verbalization: specifically to a cluster of processes that are involved in the choice of a particular linguistic expression for a CC. Subconceptualization breaks down an initial, holistic chunk into smaller chunks. These smaller chunks, however, remain conceptual in nature, and other operations are necessary to convert them into surface linguistic representations. Roughly speaking, lexicalization involves the choice of "words" that will appropriately communicate the content of CCs.

Lexicalization of a CC takes place at the point where the speaker decides that he has subconceptualized far enough. The aim of subconceptualization is to produce chunks of a size appropriate to linguistic expression, and particularly to linguistic expression that will convey neither too little nor too much information to the addressee. Too little information is, for example, provided by a summary, where subconceptualization has proceeded only to a point where lexicalization will provide the addressee with a "general idea" of the content of the whole. At the other end of the scale, we are all familiar with expositions in which too much information is conveyed, where we are told more than we want to know. One aspect of a speaker's creativity, then, is to decide exactly where in the process of subconceptualization he should stop, taking into account the needs

and interests of the addressee. It is at this point that he turns to lexicalization.

The speaker may also be influenced in such decisions by the resources his language makes available for packaging chunks of different sizes. Consider, for example, the amount of content that is packaged in an English sentence like "He hit into a double play." If our language did not provide this particular expression, we would have to subconceptualize this chunk considerably further and come up with chunks that would have to be expressed in some such way as "He hit the ball to the shortstop, who threw it to the second baseman before the runner previously on first base could reach second. The second baseman then threw the ball to the first baseman before the batter could reach first. Thus his hit caused two outs to be made." Presumably a language makes available packaging at various levels of subconceptualization according to predominant communicative needs within the culture of its speakers.

How are conceptual chunks communicated? One way to approach this question is by looking at the spatial and temporal properties of such chunks. A chunk is typically either an event ("He rubbed his left eye") or a situation ("The glue was next to the lamp"). Both events and situations have a particular locus in space and time (the difference being that an event involves some spatial change through time, whereas a situation does not). Such chunks, then, can be regarded as as-

signable to particular coordinates in both a spatial and a temporal continuum. (We omit consideration here of generic chunks, expressed in sentences like "Dogs chase cats" or "The house had two chimneys", where at least temporal particularity is absent. Genericness calls for extended discussion that would take us too far afield at this point.)

If we assume that most of the chunks a speaker wants to find linguistic expression for are events or situations, and thus have both spatial and temporal particularity, it is not surprising that language fails to provide direct labels for them. We cannot, in the course of subconceptualization, arrive at something like CC-1011, then remember that the name for this chunk is "BLURG", and communicate it by uttering that word. Particular events and situations are too numerous, and our experience of them too idiosyncratic for them to have their own particular names. The way this problem is solved is through the interpretation of many different CCs as instances of the same category. Thus the time last December when I gave my mother a Christmas present, the time when the mailman gave me a registered letter this morning, the time yesterday when the teacher gave my son a note to take home, etc. etc. are all categorizable as instances of "giving". We label the category itself UC-"GIVE" (UC standing for "universal category") and specify the choice of this category by the speaker with the notation:

27) CC-1053 C▸ UC-"GIVE"

Such a statement is to be read "CC-1053 is categorized as an instance of the category UC-"GIVE"". It should be noted that the English word "GIVE" is not the name of this category; rather, any particular CC which is so categorized can be communicated with the word "GIVE". In other words, the decision described in 27) allows us to use "GIVE" as a name for CC-1053.

The way in which a speaker decides that a particular CC can be categorized as an instance of some UC is of course a fundamental psychological question. One thing that seems clear is that some CCs are more easily categorized than others; ease of categorizability has been called "codability".[1] In a closer approximation to human mental processes, therefore, a statement like 27) ought to be qualified as valid to a certain degree, and not as an all-or-nothing decision. If the degree to which a particular CC is an instance of some UC is very high--if the CC is highly codable--then the use of the word provided by the UC will succeed quite well in conveying the content which the speaker has in mind. If, on the other hand, the content of the CC is not very well captured by assigning it to the UC, then the speaker is likely to want to add one or more modifiers to mold the content more closely to the content of the CC he has in mind. Adverbs are an obvious device by which such molding is accomplished. Thus, the speaker might decide that the content of CC-1053 is better captured in an intersection of UC-"GIVE" and UC-"GRUDGING":

28) CC-1053 C> UC-"GIVE" & UC-"GRUDGING"

in which case the eventual lexicalization will be "give grudgingly", and not simply "give".

Suppose CC-1053 is a conceptual chunk that will eventually be verbalized with the sentence:

28) Mrs. Brown gave Tommy a cookie.

We have said that the word "GIVE" is available as a label for this CC. Up to a point that is correct; there was a giving which took place. But sentence 28) contains more than the word "GIVE". What kind of conceptual information is conveyed by "MRS. BROWN", "TOMMY", and "A COOKIE"? Each of these items evidently communicates a concept that is different in nature from a CC. This other kind of concept we label a PI (for "particular individual"). The chief difference between a PI and a CC seems to have to do with temporal particularity. A CC is conceived of as occupying a specific and usually fairly limited period of time. The time period occupied by, say, Mrs. Brown is much less specific, and is not likely to be something we are very interested in when we utter a sentence like 28). In other words, while a PI may have temporal particularity in the sense of a lifespan or total time of existence, such a time period tends to be of a different order of magnitude from that occupied by a CC, and more often than not is of little relevance when the PI is communicated. Furthermore, any one PI may participate in an indeter-

minate number of different CCs. (Mrs. Brown has done many other things besides that which was reported in 28).)

Why do PIs play a necessary role in the communication of a CC? The answer may have something to do with the necessity for providing anchor points in the addressee's mind. Because of its lack of temporal particularity, the concept of a PI is a relatively stable concept, and one which is liable to enter consciousness again and again with respect to a wide variety of CCs. Thus, the only way a speaker can effectively install the content of a CC in the addressee's mind is to tie it to one or more PIs already known to the addressee. That is, the usual way of communicating information is by bringing one or more PI nodes into the addressee's consciousness, and by predicating something of these nodes. Language usually involves taking one PI (the "topic") as a starting point and either predicating something of it alone, or tying it to other PIs through a relational predicate.

In deciding to categorize a CC in a certain way, say as an instance of UC-"GIVE", a speaker simultaneously establishes a framework of PIs which are separated out from the content of the CC, and which will have to be linguistically represented in some way. In the case of UC-"GIVE" these PIs will function as agent, beneficiary, and patient (the giver, the givee, and the given). The fact that these three PIs are entailed by the choice of UC-"GIVE" is expressed as follows:

29) CC-A C> UC-"GIVE"
E>
CC-A F> VB-"GIVE" (PI-B↑AGT, PI-C↑BEN, PI-D↑PAT)

The letters A, B, C, and D in this statement are variables ranging over particular four digit numbers. For example, CC-A might be CC-1053, PI-B might be PI-1687, etc. The symbol E> is to be read "entails", and F> is to be read "is framed as". (The notation to the right of F> can be regarded as a "case frame"; hence the appropriateness of the term "framing". One might also imagine that this kind of operation involves "framing" an utterance in the sense of deciding on its basic linguistic framework.)

The statement in 29), then, says that when one has chosen to categorize a particular CC as an instance of UC-"GIVE", this decision entails that the CC will be framed as, or expressed by, the verb (VB) "GIVE" accompanied by three PIs, functioning as agent, beneficiary, and patient. Statements like that in 29) are stored in our English lexicon. This statement actually forms only part of the lexical entry for UC-"GIVE". The complete entry for this category contains a number of additional lines which state various other entailments, for example that giving involves transfer of ownership. These other aspects of lexical entri s will be discussed below.

To summarize, a CC of the appropriate size, arrived at through subconceptualization, will be subject to categorization in terms of some UC, the effect of which will be to create, by way of the lexicon, a verbal label for the CC together with a

framework of associated nouns. The framing operation, in effect, will have factored out those elements (PIs) having no significant temporal particularity, leaving a word (the VB) to which alone that temporal particularity will be assigned.

It is probably a consequence of its being left with this temporal role that the VB is likely to end up carrying a temporal marker of some kind, such as a tense and/or aspect suffix. If, for example, the CC occupies a temporal locus that precedes the locus of the speech act, the VB is likely to end up with a past tense suffix attached. This part of lexicalization we call inflection. Its implementation will be illustrated immediately below.

Our program tries to establish at the outset for each CC whether it can be categorized, on the assumption that the speaker is aiming at such categorization as a goal, and that subconceptualization takes place only when the content of the CC is such that categorization is not appropriate. Thus the first question asked of any CC is of the sort:

30) V: CAN CC-1053 BE CATEGORIZED?

If the user's answer is no, VAT goes on to ask how this CC is to be subconceptualized, as in the example given in section III. If, on the other hand, the user's answer is yes, VAT will go on to ask questions relevant to the tense/aspect properties of the CC. At present it asks first:

31) V: IS CC-1053 GENERIC?

since special considerations have to be given to CCs that do not have temporal particularity. If the answer to 31) is no, VAT presently assumes as a default option that CC-1053 has a temporal locus preceding that of the speech act. This is certainly the most probable state of affairs for most kinds of discourse. We expect later to elaborate other possibilities, which are likely to depend on adverbial and other means of establishing temporal particularity. Our program at present will, under these circumstances, add the inflectional notation "PAST" after a slash, as in:

32) CC-1053 / "PAST"

It is now time for the following exchange:

33) V: HOW IS CC-1053 CATEGORIZED?

U: GIVE

The user says that the decision has been to categorize this CC as an instance of the category UC-"GIVE". VAT then looks into the lexicon and, on the basis of the last line in 29), replaces 32) with:

```
34)  VB-"GIVE" / "PAST"
        PI-B↑AGT
        PI-C↑BEN
        PI-D↑PAT
```

Two other considerations are relevant at this point. For

one thing, VAT will want to replace the PI variables in 34) with particular four digit numbers. Our easiest recourse at present is to have VAT ask the user about each PI:

35) V: WHAT IS THE AGENT?

U: PI-1234

V: WHAT IS THE BENEFICIARY?

U: PI-1345

V: WHAT IS THE PATIENT?

U: PI-1456

whereupon VAT will replace 34) with:

36) VB-"GIVE" / "PAST"
PI-1234↑AGT
PI-1345↑BEN
PI-1456↑PAT

At least some of the answers to the questions in 35) ought, under some circumstances, to be derivable from the context. We hope gradually to teach VAT to discover such answers for itself whenever possible.

A second consideration at this point is to establish which PI is the topic. Again the easy way out is for VAT to ask the user:

37) V: WHAT IS THE TOPIC?

U: PI-1234

In English, at least, this may be the point at which functional

relations such as agent, beneficiary, and patient should be replaced by surface syntactic roles like subject, indirect object, and direct object. (In Japanese the introduction of particles like wa, ga, o, and ni would be appropriate here.) Thus, after 37) VAT may change the representation in 36) to:

38) VB-"GIVE" / "PAST"
PI-1234↑SUBJ
PI-1345↑IO
PI-1456↑DO

where IO and DO stand for "indirect object" and "direct object". Again, the identity of the topic will often be derivable from the context. For example, all other things being equal, topics have a tendency to remain constant from one clause to the next, agents are more likely to be topics than patients, and so on. Considerable empirical work will be necessary before all such factors have been sorted out.

If the codability of CC-1053 had been somewhat lower, and the modified categorization exemplified in 28) had been chosen, the representation at this stage would include an adverb (AV):

39) VB-"GIVE" / "PAST" / AV-"GRUDGING"
PI-1234↑SUBJ
PI-1345↑IO
PI-1456↑DO

The lexicalization of CC-1053, then, has involved categorization, possibly modification, inflection, and framing. The next step in verbalization is to lexicalize the several PIs which are contained in a representation like 38) or 39). We will see

that the lexicalization of a PI involves categorization, possibly modification, and inflection. Framing is for the most part restricted to the lexicalization of a CC.

V. Lexicalization of a PI

A PI is the concept of a concrete object, be it animate or inanimate, or of an abstraction which has been reified and is being treated linguistically in ways analogous to the treatment of physical objects. The surface linguistic representation of a PI may be a proper noun, a common noun, a pronoun, or nothing at all. Furthermore, by agreement processes certain features of the PI may be incorporated into the verb with which it is associated. Each language has its own idiosyncrasies in the treatment of PIs. Some, like Japanese, are especially fond of deleting the PI altogether whenever it is predictable from context. Some, of the polysynthetic type, seem to go overboard in the extent to which they incorporate features of the noun within the verb. Some make a point of adding inflectional features expressing "definiteness", plurality, and the like to the surface noun, while others seem to get along well without such expression. For illustrative purposes we will confine ourselves in this section to the main outlines of how a PI is lexicalized in English.

Much depends on whether or not the PI in question is "given"--whether it is a piece of knowledge that the speaker believes has already been brought into the addressee's consciousness in some way, prior to the uttering of the present sentence.[2] Here again we have a case where the easiest course

for VAT is to ask the user:

40) V: IS PI-1234 GIVEN?

Certainly in many cases, however, VAT can be taught to decide this for itself. If, for example, PI-1234 was mentioned in the preceding sentence the answer to 40) must be yes. If the preceding sentence was "Mrs. Brown came over from next door" and we are concerned with the lexicalization of PI-1234 within the sentence "PI-1234 gave Tommy a cookie", the givenness of PI-1234 will result in its lexicalization as "SHE". We can actually go a fair distance in establishing the givenness of a PI on this basis alone, but the question of how else givenness is established, including its introduction from knowledge external to the linguistic text altogether, will need to be raised eventually.

Let us assume first that the answer to 40) has been yes, in which case English is likely to lexicalize PI-1234 with a pronoun. This is not always the case; sometimes a PI that is given will not be pronominalized. The principal criterion here seems to be whether pronominalization will produce ambiguity, and ultimately VAT will need to decide whether ambiguity will result. For now, however, we proceed on the assumption that a PI which is given will automatically be pronominalized.

The procedure we are currently using for pronominalization in English asks first:

41) V: IS PI-1234 THE ADDRESSEE?

This question is asked first because the pronoun "YOU" does not distinguish number, and if the answer to 41) is yes it will not be necessary for VAT to do anything beyond lexicalizing PI-1234 as NN-"YOU" (NN, of course, for "noun"). If, on the other hand, the answer to 41) is no, then VAT must ask:

42) V: WHAT IS THE CARDINALITY OF PI-1234?

We assume that a PI is from one point of view the concept of a set of objects, and that the cardinality of the set is relevant in establishing expressions of singularity and plurality, among other things. Actually the distinction between one and more than one as possible answers to 42) is all that is relevant at the moment. More interesting questions do arise in this area. For example, with cardinalities up to about five there is likely to be a need for distinguishing each member of the set with a specific PI number, whereas with larger cardinalities the set is likely to be conceived of simply as containing "a number of" or "many" members.

If we assume first that the answer to 42) is one, then VAT will ask:

43) V: IS PI-1234 THE SPEAKER?

If the answer is yes, then PI-1234 is lexicalized as NN-"I". If no, then we are dealing with a third person referent and VAT

must determine its gender:

44) V: IS PI-1234 ANTHROPOMORPHIC?

This classification includes human beings, but also named animals such as pets. If the answer to 44) is no, VAT will lexicalize PI-1234 as NN-"IT". Otherwise it must find the sex of this referent:

45) V: IS PI-1234 MALE OR FEMALE?

and lexicalize it as NN-"HE" or NN-"SHE" accordingly.

If the answer to 42) was a number greater than one, VAT must decide between "WE" and "THEY", the pronouns which are explicitly plural. Essentially it must ask:

46) V: IS THE SPEAKER A MEMBER OF PI-1234?

If yes, it will produce the lexicalization NN-"WE" and if no, NN-"THEY".

There are again a variety of ways in which VAT might be able to answer questions like 41) through 46) without asking the user. The identity of speaker and addressee will have been established by providing such discourse parameters at the very beginning of the discourse; at present we use the arbitrary convention that PI-1001 is the speaker and PI-1002 the addressee. In questions 41) and 43) VAT is asking whether PI-1234 is identical to PI-1002 or PI-1001. But, depending on the context,

this identity may already have been established. As for the cardinality of PI-1234, it may have been made explicit through a numeral or in some other way. And the gender of this referent might have been established through the previous use of a sex-specific proper name, or through some other fact that has already been supplied.

Let us now turn to the possibility that PI-1234 is not given--that the answer to question 40) was no. In that case, lexicalization must be either in terms of a proper name, or through the use of a categorization and ultimately a common noun. VAT first asks:

47) DOES PI-1234 HAVE A NAME?

If yes, the user gives the name and VAT lexicalizes PI-1234 as NN-"JOHN" or the like. The real situation is not quite this simple, since a PI is likely to have more than one proper name (John, Mr. Brown, Daddy, etc.) and the choice of which, if any, among them to use will depend on various interpersonal considerations. Eventually our program should include questions relevant to such a choice.

If the answer to 47) is no, then VAT follows a procedure roughly analogous to that associated with the categorization of a CC:

48) V: HOW IS PI-1234 CATEGORIZED?

U: TEACHER

(for example). Basically, at this point, VAT will replace PI-1234 with NN-"TEACHER". At the same time it will store the statement:

49) PI-1234 <C> UC-"TEACHER"

and will look at the lexical entry for this category for whatever relevant information is stored there.

Just as a CC may be given a lexicalization that is inflected for tense and/or aspect, the lexicalization of a PI may be given inflections such as number and/or definiteness. If the lexicon shows, for example, that UC-"TEACHER" entails that PI-1234 is countable, VAT will also in this case ask about its cardinality, as in 42) above. If the answer is a number greater than one, VAT will create a representation like NN-"TEACHER" / "PLURAL". Independent of this number question, VAT will need to determine whether the use of this category in this context will enable the addressee to know what particular instance of the category is being talked about. We put this in terms of the question:

50) V: DOES UC-"TEACHER" IDENTIFY PI-1234?

If yes, VAT will add the definite article (AR) as an inflection: NN-"TEACHER" / AR-"THE". If no--that is, if the addressee is assumed not to be able to identify a previously known PI as the referent, VAT will decide between the indefinite articles AR-"A" and AR-"SOME" depending on whether the cardinality of PI-1234 is

one or greater than one. The outcome will thus be either NN-"TEACHER" / AR-"A" or NN-"TEACHER" / "PLURAL" / AR-"SOME"; that is, "a teacher" or "some teachers". We have attempted to formalize some of the contextual grounds on which VAT will be able to answer a question like 50) without asking the user, and this matter will be discussed in section VII, below.

VI. The Lexicon

In all its operations VAT must at many points make access to a store of more or less permanent lexical knowledge which we have formalized in terms of entailments of categories. The statements in the lexicon specify what we know about a particular CC or PI as a result of its being identified as an instance of a certain category. Or, to look at it from the opposite point of view, these statements say what properties a particular CC or PI must have in order to be categorized in a certain way. From the first point of view we can say that once we know that a particular CC has been categorized as an instance of UC-"GIVE", for example, the lexicon tells us a number of other things that we must know about this CC. From the second point of view we can say that the lexical entry for UC-"GIVE" tells us what we must know about a CC in order to assign it to this category. These two ways of viewing lexical entries are not in contradiction, but are different sides of the same coin.

From a psychological standpoint the lexicon approximates a description of everything that is involved in a person's interpretation of the world, at least so far as his interpretive grid is dependent on verbal categories. We are unable, of course, to focus on individual differences, but must attempt to deal with a core that is common to the speakers of a particular language. The lexicon is the heart of our program, whether we are engaged

in verbalization, translation, or parsing, and everything else depends on the success with which the lexicon has been elaborated. A separate lexicon has to be developed for each language with which the program tries to deal. In a full-fledged implementation certainly a very high proportion of the total developmental effort will have to be devoted to lexical questions.

As a simple illustration of the kind of information a lexical entry might contain, as well as of the formalism we have been using to represent such information, let us consider at least part of what it means for a particular CC to be categorized as an instance of UC-"LIFT". We will want to say that when X lifts Y, this entails that X does something which causes a change of state from Y being in one location to Y being in another location, and furthermore that the new location is above the old location. The lexical entry for UC-"LIFT", insofar as it captures this much information, is written as follows:

```
51)  CC-A  C>  UC-"LIFT"
     E>
     CC-A  F>  VB-"LIFT" (PI-B↑AGT, PI-C↑PAT)
     CC-A  S>  CJ-CAUSE (CC-D, CC-E)
     CC-D  F>  VB-ACT (PI-B)
     CC-E  S>  CJ-CONJUNCTION ((CJ-CHANGE (CC-F, CC-G)), CC-H)
     CC-F  F>  VB-AT (PI-C, PL-I)
     CC-G  F>  VB-AT (PI-C, PL-J)
     CC-H  F>  VB-ABOVE (PL-J, PL-I)
```

The first two lines are to be read, "If CC-A is categorized as an instance of UC-"LIFT", this entails..." The first line under E> then gives the case frame, saying that there will be a clause containing the verb "LIFT" accompanied by an agent (PI-B) and a

patient (PI-C). The second line under E> says that it is alternatively possible to subconceptualize CC-A in a certain way, which amounts to a paraphrase. That is, although the speaker has chosen not to subconceptualize CC-A further (presumably because the choice of UC-"LIFT" has been judged to provide the right packaging for CC-A), if he had decided to subconceptualize further he could have done it in the manner specified in this line, where two new CCs, CC-D and CC-E, are joined by CJ-CAUSE. In other words CC-D is conceived of as causing CC-E. The third line under E> says something about the content of CC-D, namely that it involves an act by PI-B. (It may be noted that the absence of quotes around ACT in VB-ACT indicates that this is not a conceptual unit that will lead to a direct surface structure representation, as will VB-"LIFT".) The fourth line under E> says that CC-E, which is caused by this act, can be subconceptualized into two conjoined elements. The first of these is a CHANGE from CC-F to CC-G, and the second is CC-H. The fifth and sixth lines under E> specify the nature of the prior and subsequent states, CC-F and CC-G. Both involve PI-C being at some location, first PL-I and then PL-J (PL standing for "particular location"). The last line elucidates CC-H, stating that the new location (PL-J) is above the old location (PL-I). Thus 51) has captured formally the several bits of knowledge about CC-A that were summarized discursively at the beginning of this paragraph.

Let us now turn to a more complicated example. This example came up initially as a result of the observation that the

Japanese verb kasu can be translated into English as either rent (out) or lend. In other words this verb is nonspecific as to whether the agent does or does not receive money for the goods or services he provides. We were interested in how a translation from Japanese into English would decide whether to use rent or lend where the Japanese had used kasu. This problem led us to consider lexical entries for several verbs involving transfers and transactions, and we arrived at a system of cross-referencing and embedding within lexical entries that captures the content of abstract notions (such as transfer and transaction) at the same time that it links entries one to another in a way that is generally useful.

We may begin by defining a transfer. We assume a category UC-TRANSFER which, since it does not contain quotation marks, is understood to be abstract and not immediately convertible into a surface structure verb. The lexical entry reads as follows:

52) CC-A C> UC-TRANSFER
E>
CC-A S> CJ-CHANGE (CC-B, CC-C)
CC-B F> VB-HAVE (PI-D, PI-E)
CC-C F> VB-HAVE (PI-F, PI-E)

Discursively, a CC-A which has been categorized as an instance of UC-TRANSFER can alternatively be subconceptualized (or paraphrased) in terms of a change from CC-B to CC-C, where the former involves PI-D "having" PI-E, and the latter involves another party, PI-F, having PI-E. In other words, a transfer involves a change in the having of some object (PI-E) from one individual

to another. The English word have of course performs a variety of semantic functions; our use of it in this formalism is meant to include at least two varieties of having--ownership, which we will label HAVE-OWN, and having the use of something, which we will call HAVE-USE. Simple HAVE, as in 52), is meant to be nonspecific as to which of these varieties of having is involved, as may be accounted for with the following two statements:

```
53)  CC-A  C>  UC-HAVE-OWN
     E>
     CC-A  C>  UC-HAVE

     CC-A  C>  UC-HAVE-USE
     E>
     CC-A  C>  UC-HAVE
```

One example of a transfer is the kind which is categorizable with UC-"GIVE", whose lexical entry can be given as follows:

```
54)  CC-A  C>  UC-"GIVE"
     E>
     CC-A  F>  VB-"GIVE" (PI-B↑AGT, ?PI-C↑BEN, PI-D↑PAT)
     CC-A  C>  UC-TRANSFER
          PI-D  =  PI-B
          PI-F  =  PI-C
          PI-E  =  PI-D
```

That is, a CC which has been categorized as an instance of UC-"GIVE" has the case frame shown in the first line under E>. The question mark before the beneficiary indicates that it is optional; one can say "Roger gave a book" without mentioning a beneficiary. The second line under E> shows that this CC can

also be categorized as an instance of UC-TRANSFER. This fact means that the CC also has the entailments listed in 52). Since the variables within each lexical entry are arbitrarily labeled A, B, C, etc., it is necessary now to state equivalences between the variables in the entry for UC-"GIVE" and those in the entry for UC-TRANSFER. These equivalences are listed, indented, in the last three lines of 54). They are to be read, "PI-D of the TRANSFER entry is equivalent to PI-B of the "GIVE" entry (the giver); PI-F of the TRANSFER entry is equivalent to PI-C of the "GIVE" entry (the givee); and PI-E of the TRANSFER entry is equivalent to PI-D of the "GIVE" entry (the given)." In this way 54) and 52) are brought into the correct alignment.

Another, more complicated kind of transfer is that involved in the category UC-"LEND":

```
55)  CC-A  C>  UC-"LEND"
     E>
     CC-A  F>  VB-"LEND" (PI-B↑AGT, ?PI-C↑BEN, PI-D↑PAT)
     CC-A  C>  UC-TRANSFER
          PI-D  =  PI-B
          PI-F  =  PI-C
          PI-E  =  PI-D
          CC-B  =  CC-E
          CC-C  =  CC-F
     CC-E  C>  UC-HAVE-USE
     CC-F  C>  UC-HAVE-USE
     VB-HAVE-OWN (PI-B, PI-D)
     CC-A  -C>  UC-TRANSACTION
```

The first seven lines of this entry are entirely parallel to the entry for UC-"GIVE" in 54). It then becomes necessary to refer to the earlier and later states, CC-B and CC-C, of the TRANSFER entry. These are equated with CC-E and CC-F of the "LEND"

entry. It is said that both of these states involve HAVE-USE. That is, when X lends an object to Y, in the earlier state X has use of the object and in the later state Y does. The next to last line says that PI-B, the agent of the lending, maintains ownership of PI-D throughout. The last line says that CC-A cannot be categorized as a transaction, as explained below. Evidently the only difference between 55) and the entry for UC-"KAS-" (i.e. kasu) in Japanese is that for the latter the last line of 55) is missing. Thus, kasu leaves it undecided whether a transaction was involved or not.

What, then, is a transaction? Essentially it is a linking of two transfers, where one of the transfers is for the purpose of the other. In buying, for example, a typical transaction, the buyer gives money to the seller so that the seller will give him some object in return. With buying, a change of ownership is involved in both transfers, but that need not be the case. With renting, for example, there is a change of ownership of the money, but only a change of use of the object. We define a transaction as follows:

```
56)  CC-A  C>  UC-TRANSACTION
     E>
     CC-A  S>  CJ-PURPOSE (CC-B, CC-C)
     CC-B  C>  UC-TRANSFER
           PI-D  =  PI-D
           PI-E  =  PI-E
           PI-F  =  PI-F
     CC-C  C>  UC-TRANSFER
           PI-F  =  PI-D
           PI-E  =  PI-G
           PI-D  =  PI-F
```

The first line under E> states that CC-A can be paraphrased in terms of CC-B and CC-C, the former being for the purpose of the latter. CC-B is a transfer in which PI-D (e.g. the buyer) transfers PI-E (e.g. money) to PI-F (e.g. the seller). CC-C is a transfer in which the roles of PI-D and PI-F (and hence their relation to the variables in 52)) are reversed. Furthermore, the object transferred (e.g. the thing bought) is a different one--here PI-G.

Besides buying and selling, another typical transaction is renting. The English word rent is ambiguous, and we will illustrate here the entry for what we call UC-"RENT-2", which is renting out (German vermieten):

```
57)  CC-A  C>  UC-"RENT-2"
     E>
     CC-A  F>  VB-"RENT" (PI-B↑AGT, ?PI-C↑BEN; ?PI-D↑MSR,
                                                PI-E↑PAT)

     CC-A  C>  UC-TRANSACTION
          PI-F  =  PI-B
          PI-D  =  PI-C
          PI-E  =  PI-D
          PI-G  =  PI-E
          CC-B  =  CC-F
          CC-C  =  CC-G
     CC-F  C>  UC-TRANSFER
          CC-B  =  CC-H
          CC-C  =  CC-I
     CC-G  C>  UC-TRANSFER
          CC-B  =  CC-J
          CC-C  =  CC-K
     PI-D  C>  UC-MEDIUM-OF-EXCHANGE
     CC-H  C>  UC-HAVE-OWN
     CC-I  C>  UC-HAVE-OWN
     CC-J  C>  UC-HAVE-USE
     CC-K  C>  UC-HAVE-USE
     VB-HAVE-OWN (PI-B, PI-E)
```

The first line under E> gives the case frame, which includes two

obligatory cases, an agent and a patient ("Bill rented (out) his lawnmower") and an optional beneficiary and measure (MSR) ("Bill rented his lawnmower to Tom for five dollars"). The second line under E> says that CC-A is a transaction; it thus conforms to 56) and it is necessary to state the equivalences between the PIs in 57) and those in 56). Below these PI equivalences it is also stated that the CC-B of the TRANSACTION definition (the transfer of money) is equivalent to CC-F of the "RENT-2" definition, while CC-C of the TRANSACTION definition (the transfer of the object) is equivalent to CC-G of "RENT-2". The two states of the first TRANSFER are named CC-H and CC-I, while the two states of the second TRANSFER are named CC-J and CC-K. It is then said that the measure, PI-D, must be something categorizable as a MEDIUM-OF-EXCHANGE--normally money, but potentially anything that would perform this function. The two states of the first TRANSFER are then both said to be instances of UC-HAVE-OWN, since the money actually changes ownership. The two states of the second transfer, on the other hand, are instances of UC-HAVE-USE, since the object does not change ownership, but only use. The last line, like the next to last line of 55), says that the agent of the renting retains ownership of the object.

It was mentioned that the lexical entry for Japanese UC-"KAS-" is the same as that for English UC-"LEND", as in 55), except that the Japanese entry lacks the last line of 55) in which it is stipulated that lending cannot be a transaction. It

can now be seen that UC-"KAS-" is compatible with both 55) and 57). We thus have a formal explanation for the fact that kasu may be translated as either lend or rent. In order to decide between the two translations, it is necessary to search the context in which this CC occurs to discover whether it is or is not a transaction. We will return to this matter in our discussion of translation in section VIII.

Lexical entries for categories whose instances are PIs are designed to elucidate the knowledge which is entailed by the assignment of a particular PI to some category. Such entries do not contain a case frame, but are otherwise similar in format to the entries for categories whose instances are CCs, as described above. As a simple example, we may note that when a PI is categorized as an instance of UC-"CAR" there is an entailment that this PI will "have" a trunk. This kind of having is different from those discussed in connection with transfers and transactions in the last section; we represent it with HAVE-AS-PART:

58) PI-A C> UC-"CAR"
E>
VB-HAVE-AS-PART (PI-A, PI-B)
PI-B C> UC-"TRUNK"

It is useful here (and elsewhere in the lexicon) to distinguish between necessary entailments and expected entailments or default options. The latter constitute knowledge that is normally entailed by the category, but not necessarily so. We in-

dicate entailments of this sort with a prefixed "E:". As an example we may note that something which has been categorized as a MEDIUM-OF-EXCHANGE (cf. 57)) is normally expected to be money, although in some circumstances it might be cowry shells or wampum:

```
59)  PI-A  C>  UC-MEDIUM-OF-EXCHANGE
     E>
     E:  PI-A  C>  UC-"MONEY"
```

A more complex example involves the categorization of a PI as an instance of UC-"BEAGLE". In this case we know that the PI is also categorizable as an instance of UC-"DOG", that we may expect that it will have a tail (although some dogs do not), that that it will bark, and that it will chase cats:

```
60)  PI-A  C>  UC-"BEAGLE"
     E>
     PI-A  C>  UC-"DOG"
     E:  VB-HAVE-AS-PART (PI-A, PI-B)
     PI-B  C>  UC-"TAIL"
     E:  VB-BARK (PI-A)
     E:  VB-CHASE (PI-A, PI-C)
     PI-C  C>  UC-"CAT"
```

It may be that E: should be expressed as a probability; that is, that there is a continuous range over which we may expect something to be entailed, with necessary entailment being one extreme. At least for practical purposes, however, it proves useful to make a three-way distinction between necessary entailments (unmarked), default expectations (E:), and a third type which we call optional entailments and mark with "O:". These last represent a lower degree of probability; they are

entailments which are neither necessary nor expected, but which are easily possible. For example, a bicycle need not have a basket and is not expected to have a basket, but it may very well have one:

```
61)  PI-A  C>  UC-"BICYCLE"
     E>
     O:  VB-HAVE-AS-PART (PI-A, PI-B)
     PI-B  C>  UC-"BASKET"
```

The distinction between necessary or expected and optional entailments is of interest when it comes to the assignment of definiteness, as discussed in the following section.

VII. Discourse Information and Readjustments

A speaker needs access to three major classes of information in order to verbalize sucessfully. First, of course, he must have an idea of what he wants to talk about: the content of the verbalization. Second, he must have access to general knowledge that is relevant, the kind of knowledge that we are attempting to characterize in the lexicon. But there is a third kind also. The speaker must keep track of knowledge having to do with the very fact that he is verbalizing: knowledge about the speech act itself, and its effect on the person his verbalization is addressed to. It is this third kind of knowledge that we are calling discourse information. We are concerned in this area with such factors as the identity and social relationship of the speaker and the addressee, the time and place of the speech act, and factors which relate the content of the discourse to what is assumed to be going on in the mind of the addressee. Sometimes, moreover, it is important to keep track of the act of verbalization as an event in itself, since the verbalization may be talked about or referred to subsequently in the discourse. Discourse information is kept by VAT in temporary storage. Unlike information in the lexicon, it is specific to and even changeable within a particular discourse rather than being potentially applicable to an unlimited number of different discourses.

Our treatment of discourse information is still rudimentary and uneven. So far as speaker and addressee are concerned, we simply enter into discourse information storage statements like the following:

62) SP-SPEAKER (PI-1001)
SP-ADDRESSEE (PI-1002)

(The prefix SP stands for "system predicate"; it is used for a variety of predicates associated with discourse information.) The program makes use of this information in various ways. For example, in deciding how to lexicalize PI-1001 and PI-1002 VAT makes use of information like that in 62) in order to answer questions like 41) and 43) in section V above.

Probably in most languages to some degree, but especially in many Asian languages, the social relationship between the speaker and addressee plays a role of some kind in verbalization. We have been interested in introducing such considerations into our verbalization procedure, and have so far concentrated on the question of how VAT should decide to categorize in Japanese a PI which in English would be categorized as an instance of UC-"GIVE". There are several categories in the Japanese lexicon, all of which conform to the definition of UC-"GIVE" in 54) above, but which differ from each other with respect to the speaker-addressee relationship. How the choice can be made is most easily illustrated in the context of a translation procedure, and we will return to this example in the section IX.

VAT does little at present with considerations of the time and place of the speech act. Statements like the following can be included with discourse information:

63) SP-HERE (PL-1357)
SP-NOW (PT-1579)

(where PL stands for "particular location" and PT for "particular time"). Whether PL-1357 and PT-1579 remain throughout the discourse or are replaced by other places and times depends on the nature of the discourse itself; sometimes there will be significant changes in these parameters and sometimes not. In any case it is possible for VAT to answer questions about tense, for example, by asking whether the time of a CC that is being verbalized is before or after, or whether it includes, the time which has been specified as NOW, such as PT-1579 in 63).

Discourse information is subject to change as the discourse proceeds. The way in which VAT presently accomplishes such changes is through readjustment processes, applied immediately after each sentence has been completely verbalized. These readjustments specify the ways in which the store of discourse information has been affected by the sentence. One of them, for example, creates a CC which is the concept of the event of producing the sentence itself, which subsequently can be treated like any other event. Everything involved in the verbalization of that sentence belongs to the content of this CC. If, for example, the speaker subsequently has reason to repeat what he

originally said, he may verbalize in exactly the same way (quote himself directly), or he may "say the same thing in different words" by making different choices in categorization and so on. The relevant information is available within the CC that represents the original verbalization.

Another readjustment has to do with the establishment of "givenness" for items communicated in the sentence. For each PI-A, for example, there will be, when the sentence has been completely verbalized, a readjustment process stateable as:

64) SP-GIVEN (PI-A)

If, for example, the sentence in question was "Mrs. Brown gave Tommy a cookie" and Mrs. Brown, Tommy, and the cookie are PI-1234, PI-1345, and PI-1456 respectively, then readjustments after the production of this sentence will create the statements:

65) SP-GIVEN (PI-1234)
SP-GIVEN (PI-1345)
SP-GIVEN (PI-1456)

If any or all of these PIs occur in the next sentence, they will be pronominalized, and it will not be necessary for VAT to ask the user a question like 40) above. Thus, the next sentence might be "He took them from her gratefully."

It is difficult to decide when statements like those in 65) should be deleted from the store of discourse information--when

givenness evaporates. After a certain period of time has elapsed in which the PI has not been talked about or otherwise kept in the addressee's consciousness, the speaker will probably no longer pronominalize it. At present we let statements like those in 65) remain only through the following sentence. Thus if PI-1234, for example, does not occur in the next sentence it will not be treated as given two sentences later, and will not be pronominalized. Not all discourse works in this way, but this device provides a useful temporary approximation.

A rather similar kind of readjustment has to do with the establishment of a relation between a UC and a PI which we call SP-IDENTIFIES. The presence of this relation eventually leads to the lexicalization of the PI with the definite article. Suppose the speaker says "I bought a bicycle yesterday." During the verbalization of this sentence VAT will have created the statement:

66) PI-1987 C> UC-"BICYCLE"

That is, PI-1987 has been categorized as an instance of UC-"BICYCLE". This statement then triggers a readjustment process which creates the discourse information:

67) SP-IDENTIFIES (UC-"BICYCLE", PI-1987)

which means that when he is presented with something that is lexicalized as an instance of UC-"BICYCLE", the addressee can be expected to know what particular instance it is (in this case

PI-1987). When, during a later sentence, VAT comes to the question:

68) V: DOES UC-"BICYCLE" IDENTIFY PI-1987?

as in 50) above, it is in a position to provide its own answer without recourse to the user. Thus it will, on its own initiative, lexicalize PI-1987 with the definite article: NN-"BICYCLE" / AR-"THE". It is in ways such as this that we are attempting to increase VAT's ability to answer its own questions.

As in the case of givenness, the question arises as to when a statement like 67) should be deleted from the store of discourse information. All that is clear now is that such statements generally last longer than SP-GIVEN statements, and for the moment we do not delete SP-IDENTIFIES statements before the end of the discourse. It is undoubtedly the case, however, that some of them should be deleted sometimes, and we will also need to deal eventually with discourses in which there are multiple instances of the same category: "the first bicycle, the second bicycle, etc."

The presence of lexical information of the type that was described at the end of section VI has an interesting and desirable effect on readjustments, specifically with respect to statements like 67). As an example, we might have a lexical entry for UC-"BICYCLE" which includes:

69) PI-A C> UC-"BICYCLE"

E>
VB-HAVE-AS-PART (PI-A, PI-B)
PI-B C> UC-"FRAME"
O: VB-HAVE-AS-PART (PI-A, PI-C)
PI-C C> UC-"BASKET"

That is, something categorized as an instance of UC-"BICYCLE" has as a necessary part something categorizable as an instance of UC-"FRAME", and also has as an optional part something categorizable as an instance of UC-"BASKET". Now, it may be noted that the second line under E , which deals with the categorization of PI-B, is a statement like that in 66) above. After a sentence like "I bought a bicycle yesterday" has been produced, this line will therefore trigger a readjustment process which creates the statement:

70) SP-IDENTIFIES (UC-"FRAME", PI-1468)

(with whatever number it is appropriate to assign to this PI). As a consequence, if PI-1468 occurs in a subsequent sentence it will be lexicalized with the definite article, as in "The frame is extra large." Thus, as is desirable, definiteness is created not only for instances of the category first mentioned, but also through entailments of that category. It should also be noted that in this context it is a little odd to say "The basket is extra large", talking about PI-C. One would be more likely to say "It has a basket which is extra large", or in some other way to introduce the basket explicitly. In other words the process just described works better for necessary parts than for optional parts of the first-mentioned object (PI-A). We therefore ex-

clude from this readjustment process PIs that have been introduced through optional entailments.

VIII. Translation

The general nature of the translation procedure was outlined in section I, and diagramed in Figure 1. To summarize again, VAT will start with a text in the source language, will reconstruct the verbalization processes which produced that text, and will then itself produce a parallel verbalization in the target language. During this last procedure it will apply syntactic processes appropriate to the target language whenever it can, but at each of those many points where it must make a choice of some kind it will look across to the source language verbalization to see what choice was made there. If possible it will equate that choice directly with a corresponding choice in the target language. If no direct correspondence is available, it will compare the lexicons of the two languages to determine what correspondences are possible, and will then search the context to decide which of them should be chosen. We will be particularly concerned in this section with illustrating a case in which such a complex choice must be made--in which the zigzag arrows in Figure 1 have considerable content. First, however, it my be useful to provide a framework by illustrating a relatively simple case where the correspondences are more direct. We will use as our first example the following brief text from Japanese:

71) Reizooko o utta. Okane ga hituyoo datta kara.
refrigerator sold money needed was because

We will want to consider some of the procedures VAT will follow in translating this sentence into English:

72) I sold the refrigerator. I needed the money.

Actually our attention in this example will focus on the first sentence, since we will later want to consider the complications that are added by changing the verb in the first sentence from utta 'sold' to kasita 'rented' or 'lent'.

Let us first review the manner in which VAT will reconstruct the original verbalization of the Japanese text. Since our eventual parsing component will follow a kind of "analysis by synthesis" procedure, we will also be suggesting here the steps of the parsing program. The only difference, and of course it is a big one, is that for the moment VAT will ask that decisions be made by the user instead of itself deriving them from the text together with its own knowledge of the world. The conversation with the user will proceed as follows:

1. V: WHAT VAT TASK DO YOU WANT PERFORMED?
2. U: VERBALIZE CC-2001
3. V: WHAT IS THE GENRE?
4. U: OPEN
5. V: CAN CC-2001 BE CATEGORIZED?
6. U: NO

7. V: HOW IS CC-2001 SUBCONCEPTUALIZED?

8. U: REASON (CC-2002, CC-2003)

As explained for example 9) in section II, and with the proper insertion of periods, VAT's representation now is:

```
CC-2003
CJ-"."
CC-2002
CJ-"KARA"
CJ-"."
```

9. V: CAN CC-2003 BE CATEGORIZED?

10. U: YES

11. V: IS CC-2003 GENERIC?

12. U: NO

VAT's representation, as explained for example 32) in section IV, now includes:

```
CC-2003 / "PAST"
...
```

13. V: HOW IS CC-2003 CATEGORIZED?

14. U: UR-

VAT finds UC-"UR-" in the Japanese lexicon. The first three lines of this entry are:

```
CC-A  C>  UC-"UR-"
E>
CC-A  F>  VB-"UR-" (PI-B↑AGT, ?PI-C↑BEN, ?PI-D↑MSR,
                                          PI-E↑PAT)
...
```

As in example 34) in section IV, VAT creates the representation:

```
VB-"UR-" / "PAST"
   PI-B↑AGT
   ?PI-C↑BEN
   ?PI-D↑MSR
   PI-E↑PAT
CJ-"."
CC-2002
CJ-"KARA"
CJ-"."
```

Since the beneficiary and measure PIs are optional, VAT next asks:

15. V: IS THE BENEFICIARY EXPLICIT?

16. U: NO

17. V: IS THE MEASURE EXPLICIT?

18. U: NO

The next two questions are:

19: V: WHAT IS THE AGENT?

20: U: PI-2001

21: V: WHAT IS THE PATIENT?

22: U: PI-2003

VAT now has the following representation (cf. 36) above):

```
VB-"UR-" / "PAST"
   PI-2001↑AGT
   PI-2003↑PAT
CJ-"."
CC-2002
CJ-"KARA"
CJ-"."
```

VAT next asks:

23. V: WHAT IS THE TOPIC?

24. U: PI-2001

whereupon for Japanese it presently creates the structure:

```
VB-"UR-" / "PAST"
   PI-2001 / "GA"
   PI-2003 / "O"
CJ-"."
CC-2002
CJ-"KARA"
CJ-"."
```

VAT is now at a point where it can lexicalize PI-2001 and PI-2003. Beginning with PI-2001, it might ask first:

25. V: IS PI-2001 GIVEN?

26. U: YES

In fact, however, we assume that the speaker (and addressee) are automatically given, so that VAT contains a general entailment to the effect that:

```
SP-SPEAKER (PI-A)
E>
SP-GIVEN (PI-A)
```

Since by convention PI-2001 is the speaker, the following is already stored as discourse information:

```
SP-GIVEN (PI-2001)
```

Thus VAT was able to give an affirmative answer to question 25

above without asking the user. Pronominalization in Japanese is a complex matter, depending in part on social relationships, and we have not as yet constructed a procedure to introduce the correct pronoun for a PI that is given. We have, however, taken advantage of the simple fact that given PIs are very often deleted, with no surface representation at all. In the present example, and in many others, the simple deletion of such a PI produces the correct result, so that an affirmative answer to question 25 leads to the representation:

```
VB-"UR-" / "PAST"
   PI-2003 / "O"
CJ-"."
CC-2002
CJ-"KARA"
CJ-"."
```

VAT now turns its attention to PI-2003:

27. V: IS PI-2003 GIVEN?

28. U: NO

29. V: DOES PI-2003 HAVE A NAME?

30. U: NO

31. V: HOW IS PI-2003 CATEGORIZED?

32. U: REIZOOKO

(We omit here considerations of cardinality.) The representation now is:

```
VB-"UR-" / "PAST"
```

```
   NN-"REIZOOKO" / "O"
CJ-"."
CC-2002
CJ-"KARA"
CJ-"."
```

The first three lines of the above are actually as far as we go at the present time in the surface representation of a sentence. We try to include in such a representation everything that is needed to arrive at a correct linear sequence of words. In this case the combination VB-"UR-" / "PAST" will yield the surface word utta, which will be placed in sentence-final position (followed by the period). That leaves reizooko o as the first words in the sentence.

VAT would next ask about CC-2002, but we will not carry the verbalization process further here. We are interested in how just this much of the text will be translated into English. By and large VAT will ask the same questions it asked in the course of the Japanese verbalization. It will look for the answers in the answers that were given there, and when possible will apply corresponding answers in English. Along the way, whenever appropriate, it will apply syntactic processes that are called for by the structure of English. The translation, then, begins with the same question that began the verbalization in Japanese:

V: WHAT VAT TASK DO YOU WANT PERFORMED?

The answer given in line 2 above was VERBALIZE CC-2001. The English translation must use its own four digit numbers; in

what follows we will simply substitute the English digit "1" for the Japanese digit "2":

U: VERBALIZE CC-1001

Of course here as elsewhere this question is not actually asked of the user, but is answered internally by VAT. The next questions exactly parallel lines 3-8 above:

V: WHAT IS THE GENRE?

U: OPEN

V: CAN CC-1001 BE CATEGORIZED?

U: NO

V: HOW IS CC-1001 SUBCONCEPTUALIZED?

U: REASON (CC-1002, CC-1003)

We assume that English would not in this case use the word because, but simply juxtapose the two sentences, as in example 8) in section II. Thus the representation now is:

CC-1003
CJ-"."
CC-1002
CJ-"."

Lines 9-13 of the Japanese verbalization have a direct correspondence:

V: CAN CC-1003 BE CATEGORIZED?

U: YES

V: IS CC-1003 GENERIC?

U: NO.

V: HOW IS CC-1003 CATEGORIZED?

At this point the Japanese answer was UR-. That is, the categorization was in terms of the Japanese category UC-"UR-". It is necessary to find an English category that corresponds. The procedure at this point is to look first in a stored list of bilingual category equivalences which we call interlingua. The entries in interlingua are of the following sort:

Japanese	English
⋮	⋮
UR-	SELL
⋮	⋮
HON	BOOK
⋮	⋮

That is, the list contains pairs of categories, where the members of each pair are assumed to categorize what is, for all practical purposes, identical content. The assumption is that if a CC can be categorized as an instance of UC-"UR-" in Japanese it can also be categorized as an instance of UC-"SELL" in English, and vice versa. Similarly, Japanese UC-"HON" and English UC-"BOOK" are equivalent categories. As a general strategy we expect that pairs will gradually be removed from interlingua as differences between the paired categories are discovered. Linguistic research has not yet progressed to the point that we can say with complete certainty that any two categories from two different languages embrace exactly the same content. At the

outset, however, it is useful at least to pretend that UC-"UR-" and UC-"SELL" are equivalent, and probably there are at least some pairs in interlingua that will remain viable for some time.

The present example was chosen because the answer to the last question above can be found in interlingua. Later we will consider a case where it cannot. At this point VAT answers its own question with:

U: SELL

then looks at the lexical entry for UC-"SELL" (which we assume does not differ from that for UC-"UR-"), and creates the representation:

```
VB-"SELL" / "PAST"
   PI-B↑AGT
   ?PI-C↑BEN
   ?PI-D↑MSR
   PI-E↑PAT
CJ-"."
CC-1002
CJ-"."
```

The questions and answers which parallel lines 15-22 of the Japanese verbalization are straightforward:

V: IS THE BENEFICIARY EXPLICIT?

U: NO

V: IS THE MEASURE EXPLICIT?

U: NO

V: WHAT IS THE AGENT?

U: PI-1001

V: WHAT IS THE PATIENT?

U: PI-1003

The representation now is:

```
VB-"SELL" / "PAST"
   PI-1001↑AGT
   PI-1003↑PAT
CJ-"."
CC-1002
CJ-"."
```

The next exchange is:

V: WHAT IS THE TOPIC?

U: PI-1001

which creates the representation:

```
VB-"SELL" / "PAST"
   PI-1001↑SUBJ
   PI-1003↑DO
CJ-"."
CC-1002
CJ-"."
```

With the lexicalization of PI-1001 the procedure is different in English, since this item cannot simply be deleted as in the Japanese. We follow the questions illustrated in examples 40) through 43) in section V:

V: IS PI-1001 GIVEN?

U: YES

V: IS PI-1001 THE ADDRESSEE?

U: NO

V: WHAT IS THE CARDINALITY OF PI-1001?

U: 1

V: IS PI-1001 THE SPEAKER?

U: YES

Thus the representation now is:

```
VB-"SELL" / "PAST"
   NN-"I"↑SUBJ
   PI-1003↑DO
CJ-"."
CC-1002
CJ-"."
```

Now comes the lexicalization of the direct object, PI-1003. The initial questions parallel lines 27-31 of the Japanese verbalization:

V: IS PI-1003 GIVEN:

U: NO

V: DOES PI-1003 HAVE A NAME?

U: NO

V: HOW IS PI-1003 CATEGORIZED?

The Japanese answer was REIZOOKO. VAT will now look in interlingua to see whether that item is there, and we assume that it will be found paired with English REFRIGERATOR. Although Japanese was able to terminate the verbalization of PI-2003 at this point, English must ask the question introduced in example 50) of section V:

V: DOES CC-"REFRIGERATOR" IDENTIFY PI-1003?

The answer depends on the context, but let us assume that it is yes. The representation now is:

```
VB-"SELL" / "PAST"
   NN-"I"↑SUBJ
   NN-"REFRIGERATOR" / AR-"THE"↑DO
CJ-"."
CC-1002
CJ-"."
```

We now have the kind of representation of the first sentence that is our current goal. Normal English word order will put the subject first, the verb second, and the direct object last to yield the final representation "I sold the refrigerator" of 72).

The above example was chosen to illustrate a maximally simple case of translation: one in which, in particular, the answers to all questions about cross-language categorization could be found in interlingua. The interesting cases, however, are those in which interlingua does not provide all the answers. It is in these cases that the zigzag arrows of Figure 1 in section I must be further elaborated. The general method of elaboration is suggested in Figure 4. Assume that we are producing a verbalization in the target language and, coming down from the upper righthand corner, we arrive at a point where a CC or PI needs to be categorized. Following arrow 1, we look across to the source language verbalization to find that the corresponding CC or PI was categorized in a certain way, let us say as an instance of category A. We look next at interlingua (arrow 2).

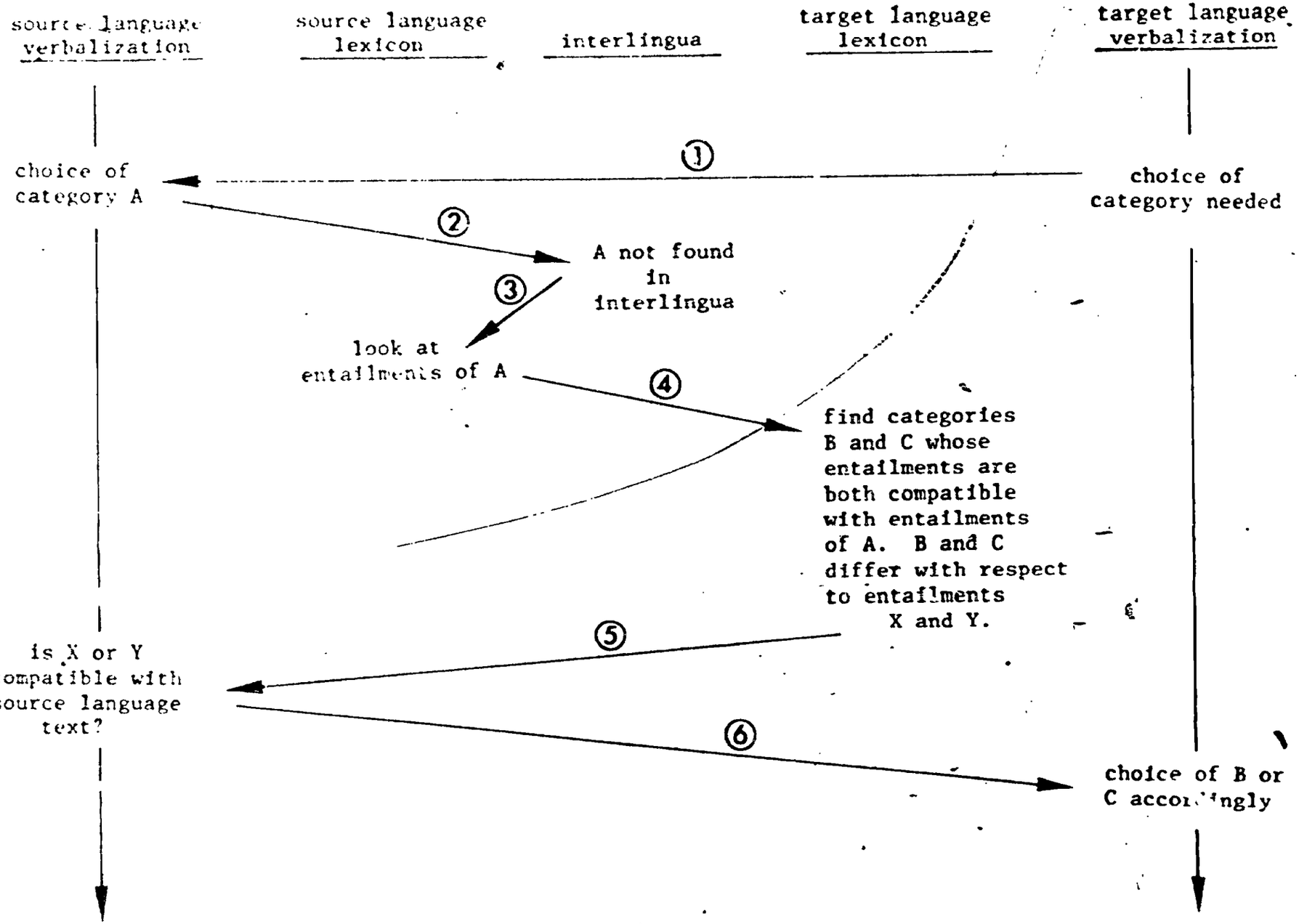
source language verbalization
source language lexicon
interlingua
target language lexicon
target language verbalization
choice of category A
choice of category needed
A not found in interlingua
look at entailments of A
find categories B and C whose entailments are both compatible with entailments of A. B and C differ with respect to entailments X and Y.
is X or Y compatible with source language text?
choice of B or C accordingly
1
2
3
4
5
6

Figure 4

If A were there, we would take the target language category paired with it (such as SELL and REFRIGERATOR in the example above), introduce it into the target language verbalization, and proceed. Now, however, we are considering those cases in which A is not found in interlingua. The next step, following arrow 3, is to look at the entailments of A in the source language lexicon. We next follow arrow 4 to search the target language lexicon for entries whose entailments are compatible with those of A. (This search procedure is likely to present challenging problems when the source language lexicon reaches any interesting size. It is, however, facilitated by the presence of abstract features like TRANSFER and TRANSACTION which can be used to limit the domain of search.) Suppose that we find two entries in the target language lexicon, B and C, both of whose entailments are compatible with the entailments of A. We then look to see how the entailments of B and C differ and find, let us say, that B contains entailment(s) X while C contains entailment(s) Y. We then follow arrow 5 back to the source language verbalization, hoping to find something in it that will allow us to choose between X and Y. (Again there are challenging problems in searching the source language text for the answer, problems that we have hardly begun to deal with.) Let us now assume that we find something in the source language text that is compatible with X but not with Y. We are then able to choose B as the correct target language category. We introduce that category into the target language verbalization via arrow

6 and proceed. In those cases where the choice between X and Y (and hence between B and C) cannot be made--where the source language text does not provide the answer--VAT must resort to asking the user for the correct categorization.

We will illustrate this procedure with the brief Japanese text:

73) Reizooko o kasita. Okane ga hituyoo datta kara.
refrigerator rented money needed was because

We will want VAT to translate these two sentences into English:

74) I rented the refrigerator. I needed the money.

We are not concerned in this example with the fact that the first English sentence is ambiguous between rented (to someone) and rented (from someone), but with the fact that the first Japanese sentence is ambiguous between rented and lent. In both cases, it seems, the second sentence serves to disambiguate. What we are interested in now is the fact that VAT must somehow choose between RENT and LEND as the proper correspondent for Japanese KAS-.

We can assume that most of the verbalization in both languages proceeds along the lines already exemplified, since 71) and 73) are minimally different. Imagine, then, that we have arrived at the point in the English verbalization where the question is:

V: HOW IS CC-1003 CATEGORIZED?

We are now in the upper right of Figure 4, and we follow arrow 1 to find that the corresponding CC in the Japanese verbalization was categorized in terms of UC-"KAS-". We then follow arrow 2, and find that KAS- is not in interlingua. We look next via arrow 3 at the entailments of UC-"KAS-" and find that they are as specified in example 55), section VI above, but without the last line of that example:

```
75)  CC-A  C>  UC-"KAS-"
     E>
     CC-A  F>  VB-"KAS-" (PI-B↑AGT, ?PI-C↑BEN, PI-D↑PAT)
     CC-A  C>  UC-TRANSFER
          PI-D  =  PI-B
          PI-F  =  PI-C
          PI-E  =  PI-D
          CC-B  =  CC-E
          CC-C  =  CC-F
     CC-E  C>  UC-HAVE-USE
     CC-F  C>  UC-HAVE-USE
     VB-HAVE-OWN (PI-B, PI-D)
```

Substituting four digit numbers for the variables, we obtain:

```
76)  CC-2003  C>  UC-"KAS-"
     E>
     CC-2003  F>  VB-"KAS-" (PI-2001↑AGT, ?PI-2902↑BEN, PI-2003
                                                          ↑PAT)
     CC-2003  C>  UC-TRANSFER
          PI-D  =  PI-2001
          PI-F  =  PI-2902
          PI-E  =  PI-2003
          CC-B  =  CC-2905
          CC-C  =  CC-2906
     CC-2905  C>  UC-HAVE-USE
     CC-2906  C>  UC-HAVE-USE
     VB-HAVE-OWN (PI-2001, PI-2003)
```

(PI-2902, CC-2905, and CC-2906 have been inserted here as arbi-

trary numbers. It is quite possible, however, that these are items which show up explicitly elsewhere in the Japanese verbalization. For example, PI-2902, the one who receives the refrigerator, might well be mentioned elsewhere in the text.)

Since CC-2003 involves a transfer, VAT must also assign numbers within the definition of UC-TRANSFER, given in section VI above as example 52):

```
77)  CC-2003  C>  UC-TRANSFER
     E>
     CC-2003  S>  CJ-CHANGE (CC-2905, CC-2906)
     CC-2905  F>  VB-HAVE (PI-2001, PI-2003)
     CC-2906  F>  VB-HAVE (PI-2902, PI-2003)
```

Thus there is a change from the renter or lender (PI-2001) having the object (PI-2003) to the rentee or borrower (PI-2902) having it. The last three lines of 76) made it clear that this was not a change in ownership but only a change in use, and that PI-2001 retains ownership throughout.

Following arrow 4, we carry these entailments across to the English lexicon and search for entries whose entailments are compatible with 76). Compatibility means that these entries will contain what is in 76), but may also contain more. Let us say that we find two such entries, one for the category UC-"LEND', which was given in 55) above, and one for UC-"RENT-2", which was given in 57).

The next step is to isolate the differences between UC-"LEND" and UC-"RENT-2". UC-"LEND", as mentioned, differs from

75) in containing an additional final line:

78) CC-A -C> UC-TRANSACTION

That is, CC-A cannot be categorized as a transaction. UC-"RENT-2", on the other hand, contains the statement:

79) CC-A C> UC-TRANSACTION

At one level of abstraction the question which must be answered, therefore, is whether CC-1003 is or is not a transaction. Informally, this is a matter of whether PI-2001, the renter or lender, did or did not receive money in exchange for the transfer of use of the object.

The following digits can be inserted for the variables in the lexical entry for UC-"RENT-2":

```
80)  CC-1003  C>  UC-"RENT-2"
     E>
     CC-1003  F>  UC-"RENT" (PI-1001↑AGT, ?PI-1901↑BEN,
       ?PI-1902↑MSR, PI-1003↑PAT)
     CC-1003  C>  UC-TRANSACTION
          PI-F  =  PI-1001
          PI-D  =  PI-1901
          PI-E  =  PI-1902
          PI-G  =  PI-1003
          CC-B  =  CC-1901
          CC-C  =  CC-1902
     CC-1901  C>  UC-TRANSFER
          CC-B  =  CC-1903
          CC-C  =  CC-1904
     CC-1902  C>  UC-TRANSFER
          CC-B  =  CC-1905
          CC-C  =  CC-1906
     PI-1902  C>  UC-MEDIUM-OF-EXCHANGE
     CC-1903  C>  UC-HAVE-OWN
     CC-1904  C>  UC-HAVE-OWN
     CC-1905  C>  UC-HAVE-USE
     CC-1906  C>  UC-HAVE-USE
```

VB-HAVE-OWN (PI-1001, PI-1003)

What all this says is that the categorization of CC-1003 as an instance of UC-"RENT-2" involves a number of things. First, there must be a person who does the renting out (PI-1001), a person who receives the rented object (PI-1901), the money that is paid in rent (PI-1902), and the rented object itself (PI-1003). Furthermore, CC-1003 is said to be a transaction, and certain equivalences are stated between the RENT-2 definition and the TRANSACTION definition. VAT must therefore ass in these particular PI and CC numbers within the definition of UC-TRANSACTION which was given as example 56) in section VI above:

```
81)  CC-1003  C>  UC-TRANSACTION
     E>
     CC-1003  S>  CJ-PURPOSE (CC-1901, CC-1902)
     CC-1901  C>  UC-TRANSFER
          PI-D  -  PI-1901
          PI-E  -  PI-1902
          PI-F  -  PI-1001
     CC-1902  C>  UC-TRANSFER
          PI-F  -  PI-1901
          PI-E  -  PI-1003
          PI-D  -  PI-1001
```

This says that CC-1003 can be paraphrased as two transfers, CC-1901 and CC-1902, the first of which was for the purpose of the second. (CC-1901 is the transfer of money, and CC-1902 the transfer of the rented object.) VAT must, therefore, look also at the definition of UC-TRANSFER, given in section VI above as example 52), and introduce again the proper PI and CC numbers for each of these particular transfers. The first of them will be represented as:

82) CC-1901 C> UC-TRANSFER
E>
CC-1901 S> CJ-CHANGE (CC-1903, CC-1904)
CC-1903 F> VB-HAVE (PI-1901, PI-1902)
CC-1904 F> VB-HAVE (PI-1001, PI-1902)

That is, the first transfer involves a change from CC-1903 to CC-1904. In CC-1903 the rentee (PI-1901) has the money (PI-1902), and in CC-1904 the renter (PI-1001) has it. The second transfer is represented as:

83) CC-1902 C> UC-TRANSFER
E>
CC-1902 S> CJ-CHANGE (CC-1905, CC-1906)
CC-1905 F> VB-HAVE (PI-1001, PI-1003)
CC-1906 F> VB-HAVE (PI-1901, PI-1003)

Here there is a change from CC-1905 to CC-1906. In CC-1905 the renter (PI-1001) has the object to be rented (PI-1003), and in CC-1906 the rentee (PI-1901) has it.

In 80) it is also stated that PI-1902 can be categorized as an instance of MEDIUM-OF-EXCHANGE, in all probability therefore an instance of UC-"MONEY" (see example 59) in section VI above). Furthermore it is stated that the change in the having of the money (from CC-1903 to CC-1904) involves a change in ownership, whereas the change in the having of the rented object (from CC-1905 to CC-1906) involves a change in use. Finally, it is stated that the renter (PI-1001) retains ownership of the rented object throughout.

What VAT wants to find out, then, is whether these things that must be true if CC-1003 is to be an instance of UC-"RENT-2"

are indeed true, or whether the bottom line in the entailments of UC-"LEND", example 78), is fulfilled instead. VAT tries to decide this by following arrow 5 to the verbalization of the Japanese text. Of course there are many ways in which the answer might appear in that verbalization, if it appears at all. If VAT is unsuccessful in its search it will have to ask the user directly:

84) V: IS CC-1003 CATEGORIZED AS LEND OR RENT?

In 73), however, we have made things easy by supplying a context which ought to decide the question. It will be remembered that the second sentence in 73) expresses CC-2002, which is the REASON for CC-2003, or what is expressed in the first sentence. Now, CC-2002 is categorized in the Japanese as an instance of UC-"HITUYOO DA", which means something like "be needed". Let us assume that the Japanese lexicon contains an entry for this category which includes the following:

```
85)  CC-A  C>  UC-"HITUYOO DA"
     E>
     CC-A  F>  VB-"HITUYOO DA" (PI-B↑BEN, PI-C↑PAT)
     CC-A  F>  VB-WANT (PI-B, CC-D)
     CC-D  F>  VB-HAVE (PI-B, PI-C)
     ...
```

The case frame immediately under the E> identifies PI-B as the beneficiary, the person who needs something, while the thing needed is labeled PI-C. The second line under the E> says that an alternative framing is possible in terms of an abstract verb WANT, wherein PI-B wants CC-D, and CC-D is then characterized in

terms of PI-B having PI-C. In other words, when one needs something, one wants to have it. (If this is not always true, at least it is the expected entailment.)

If 85) is going to provide an answer to 84), there must also be a general principle of some kind which relates what is entailed by CC-2002 to what is entailed by CC-2003. This general principle can be stated as follows:

```
86)  CC-A  F>  VB-WANT (PI-B, CC-C)
     CC-D  F>  VB-"E" (PI-B↑AGT)
     CJ-REASON (CC-A, CC-D)
     E>
     CC-D  E>  CC-C
```

The first line says that PI-B wants CC-C. The second line says that PI-B does something. The third line says that his wanting CC-C is the reason he does something. All of this together is then said to entail that his doing something entails what he wants, or CC-C. In other words, if one wants something and does something because of that, then what one does must entail what one wants.

During the verbalization of CC-2002 as part of the verbalization of the Japanese text, VAT will have recorded the fact that CC-2002 was categorized as an instance of UC-"HITUYOO DA", and will have entered the following statements in accordance with 85):

```
87)  CC-2002  C>  UC-"HITUYOO DA"
     E>
     CC-2002  F>  VB-"HITUYOO DA" (PI-2001↑BEN, PI-2902↑PAT)
```

CC-2002 F> VB-WANT (PI-2001, CC-2904)
CC-2904 F> VB-HAVE (PI-2001, PI-2902)

At this point VAT also has all the particulars needed for principle 86), which can be filled out as follows:

88) CC-2002 F> VB-WANT (PI-2001, CC-2904)
CC-2003 F> VB-"KAS-" (PI-2001 AGT)
CJ-REASON (CC-2002, CC-2003)
E>
CC-2003 E> CC-2904

The first line of 88) was obtained from 87). The second line was obtained from 76). The third line comes from line 8 of the Japanese verbalization set forth at the beginning of this section. What we are interested in now is the last line of 88), which says in effect that CC-2003 is categorized in such a way that CC-2904 is true, and looking back to 87) we see that CC-2904 involves PI-2001 having PI-2902, or the agent of kasu having okane 'money'. Making the necessary correspondences in English, this means that CC-1003 must be categorized in such a way that CC-1904 is true, where:

89) CC-1904 F> VB-HAVE (PI-1001, PI-1902)

This is exactly what VAT finds as the last line of 82). Since 82) is entailed by UC-"RENT-2" but not by UC-"LEND", the question in 84) has been answered, and the arrow labeled 6 in Figure 4 carries back the choice of UC-"RENT-2" into the English verbalization, which then proceeds as it did in the translation illustrated earlier.

By this complex process involving comparisons of entailments within and across languages, as well as the general principle stated in 86), VAT has been able to make the correct choice. So long as the answer to 84) was derivable from something discoverable within the Japanese verbalization, VAT could in principle succeed. It is clear, however, that the route to the answer could be extremely complex, involving chains of entailments of unforeseeable length. There is no doubt that such procedures are necessary to answer such questions, and that they present an extraordinary challenge to our techniques for information storage and search.

IX. Miscellaneous Problems in Translation

Since we have spent considerable time looking into various specific translation problems beyond those illustrated above, we present here a few additional examples of the sorts of things that will have to be taken into account during the implementation of machine translation along the lines suggested above. Two of these examples will, like those in the last section, involve the choice of a category in the target language when that choice is not directly provided by interlingua. One has to do with the translation of Japanese osieru into English; the other, the translation of English give into Japanese. A third example will illustrate the kind of problem that arises at the stage of subconceptualization and sentence formation.

The following three sentences illustrate three possible English translations of the Japanese verb osieru:

90) Gaido wa Kookyo ga doko ni aru ka osiete kuremasita.
guide Imperial Palace where is showed

soko kara tookyoo tawaa e ikimasita.
there from Tokyo tower to went

The guide showed us where the Imperial Palace was.

From there we went to the Tokyo Tower.

91) Gaido wa Kookyo ga doko no aru ka osiete kuremasita
guide Imperial Palace where is told

ga watasitati ga soko e itta toki ni moo simatte
but we there to went when already closed

imasita.
was

The guide told us where the Imperial Palace was, but when we got there it was already closed.

92) Kimatu siken no tame ni sensei wa
semester-final exam of for the purpose teacher

Kookyo ga doko ni aru ka osiete kudasaimasita.
Imperial Palace where is taught

For the final exam the teacher taught us where the Imperial Palace was.

Each of these examples contains the phrase:

93) Kookyo ga doko ni aru ka osiete

which is translated in three different ways, determined by the context:

in 90): show where the Imperial Palace is

in 91): tell where the Imperial Palace is

in 92): teach where the Imperial Palace is

The difference is localized in the translation of osiete, a participial form of the verb osieru. This verb may be translated into English as show, tell, or teach according to the context, and the problem is to identify what the determining factors are.

The Japanese category UC-"OSIE-" as well as the English categories UC-"SHOW", UC-"TELL", and UC-"TEACH" are all included within the ~~more abstract category~~ UC-COMMUNICATION, which can be defined as follows:

94) CC-A C> UC-COMMUNICATION
E>
CC-A F> VB-INTEND (PI-B, CC-C)
CC-C S> CJ-CAUSE (CC-D, CC-E)
CC-D F> VB-ACT (PI-B)
CC-E S> CJ-CHANGE (CC-F, CC-G)
CC-F F> -VB-KNOW (PI-H, CC-I)
CC-G F> VB-KNOW (PI-H, CC-I)

That is, for a CC to be categorized as an instance of UC-COMMUNICATION entails that someone (PI-B) intends something (CC-C), and that what he intends is that CC-D will cause CC-E. CC-D is some act that PI-B performs, and CC-E, caused by that act, is a change from state CC-F to state CC-G. CC-F is a state in which another person (PI-H) does not know something (CC-I), and CC-G is a state in which that person does know it.

Subcategories of UC-COMMUNICATION may differ as to the nature of the act (CC-D) performed by the communicator, as to the kind of knowing that results (e.g. whether it is retained in surface or deep memory), and in other ways such as the authoritativeness of the communicator with respect to what is communicated (CC-I). The Japanese category UC-"OSIE-", for example, is less specific as to the act performed by the communicator; apparently he can do almost anything that will have a communicative function. UC-"TELL", on the other hand, entails a

verbal act, UC-"SHOW" an act which directs the other persons visual attention to CC-I, and UC-"TEACH" an act which is didactic in nature. It is difficult to delimit the acts which qualify as teaching, but evidently they must have an instructional quality which is not necessary for UC-"OSIE-". UC-"TEACH" may also be unique in requiring that the knowing (CC-G) be deep or long-term knowing, at least in the intention of PI-B.[3] Japanese UC-"OSIE-" may, for its part, require that PI-B be authoritative with respect to the content of what is being communicated (CC-I).

But how is it, for example, that the context in 90) restricts the translation of "OSIE-" to "SHOW"? The second sentence in 90) says that we went from <u>there</u> (<u>soko</u>), whose referent is the location of the Imperial Palace. Thus, at the time of the communicative event, we must have been at the Imperial Palace. Now, there is evidently a general principle, like 86) in the last section, which says that a verbal act is not used to communicate where something is when the beneficiary of the act is already at that place. There is evidently no such restriction on directing visual attention to where it is, hence UC-"SHOW" is preferred to UC-"TELL". Since there is nothing in the context of 90) to suggest that teaching methods were involved, UC-"SHOW" is left as the only candidate.

In 91) the situation is otherwise. The second clause makes it clear through the phrase translated "when we got there" that we were <u>not</u> at the Imperial Palace at the time of the communi-

cative act. Another general principle says that visual attention can be directed only at things within visual range. Thus UC-"SHOW" is in this case ruled out, as is UC-"TEACH" again because of the absence of didactic context. UC-"TELL" is thus the choice here.

In 92) the didactic context is evident. The Japanese words kimatu, siken, and sensei all belong within the semantic field of teaching, a fact to be noted in the lexical entry for each of them. Hence the English category UC-"TEACH", obviously a member of the same semantic field, will be the choice here. Probably we should also take account of the fact that the idiomatic verb at the end of this sentence, literally 'gave', reinforces the superior relationship of the communicator: in this case, the fact that he is authoritative with respect to what is being communicated.

The point of this example of the translation of osieru is to emphasize the complexity of the criteria which may have to be invoked to decide between possible translations. Here we have seen a link between different kinds of communicative acts and the location of the recipient of the communication, information on the latter being derivable from information about the movement of the recipient to or from the place of communication, together with temporal information. It is also of interest that this example, like the second example in section VIII, led us to recognize certain general principles: that one does not com-

municate verbally about where something is when the addressee is already there, for example, and the obvious principle that one does not call visual attention to something that is not visible. Detailed implementation of this kind of translation research will undoubtedly lead to the recognition of a number of such principles.

The word kudasaimasita in 92) leads us to a different kind of complication, that involved in the need to pay special attention in Japanese verbalization to the social relationship existing between the speaker and various other persons. Although we are changing the direction of translation here, it is of some interest to consider questions that arise in translating the English category "UC-"GIVE" into Japanese. We may assume that UC-"GIVE" has the entailments listed in example 54), section VI above, and that furthermore the categories underlying all the Japanese verbs to be mentioned share these same entailments. Each Japanese category, however, has additional entailments of its own, and it is the nature of these additional entailments that we are interested in.[4]

The verb kurera, for example, is used to express instances of a category whose entailments include those of UC-"GIVE" plus the following (where P1-B is the agent and P1-C the beneficiary of the giving):

95) CC-A C▸ UC-"KURE-"
E▸
...

VB-CLOSE-TO-SPEAKER (PI-C)
VB-CLOSER-TO-SPEAKER-THAN (PI-C, PI-B)
-VB-HIGHER-THAN (PI-B, PI-C)

That is, UC-"KURE-" is the category chosen if the beneficiary of the giving is socially close to the speaker, closer to the speaker than the agent of the giving, and the agent is not socially higher than the beneficiary. In translating texts where such information is relevant, VAT will either have to store a network of social relations linking all the relevant individuals, a network which may in part be derivable from the text, or it will have to ask the user questions like:

96) V: IS PI-2849 SOCIALLY CLOSE to PI-2001?
V: IS PI-2849 SOCIALLY CLOSER TO PI-2001 THAN PI-2365?
V: IS PI-2365 SOCIALLY HIGHER THAN PI-2849?

The verb kudasaru, whose idiomatic function appeared in 92), is used to express instances of a category whose entailments are as follows:

97) CC-A C> UC-"KUDASAR-"
E>
...
VB-CLOSE-TO-SPEAKER (PI-C)
VB-CLOSER-TO-SPEAKER-THAN (PI-C, PI-B)
VB-HIGHER-THAN (PI-B, PI-C)

In other words, the entailments of UC-"KUDASAR-" are the same as those of UC-"KURE-" except that the agent of the giving is socially higher than the beneficiary. (It was the exalted position of sensei, the teacher, in 92) that led to the use of

kudasaimasita in that sentence.)

Another possibility is the verb yaru:

98) CC-A C> UC-"YAR-"
E>
...
{-VB-CLOSE-TO-SPEAKER (PI-C)
{-VB-CLOSER-TO-SPEAKER-THAN (PI-C, PI-B)}
VB-HIGHER-THAN (PI-B, PI-C)
-VB-RESPECTED (PI-C)

The braces indicate a disjunction. Thus one of the ways in which this category differs from the last two is in the beneficiary of the giving not being socially close to the speaker, or else in his not being closer to the speaker than the agent of the giving. As in 97) the agent is socially higher than the beneficiary. Furthermore, as stated in the last line, the beneficiary is not being treated respectfully by the speaker.

The verb ageru is like yaru, except that the agent of the giving is not socially higher than the beneficiary:

99) CC-A C> UC-"AGE-"
E>
...
{-VB-CLOSE-TO-SPEAKER (PI-C)
{-VB-CLOSER-TO-SPEAKER-THAN (PI-C, PI-B)}
-VB-HIGHER-THAN (PI-B, PI-C)
-VB-RESPECTED (PI-C)

The last verb that we will consider here is sasiageru:

100) CC-A C> UC-"SASIAGE-"
E>
...
VB-CLOSE-TO-SPEAKER (PI-B)

VB-HIGHER-THAN (PI-C. PI-B)
VB-RESPECTED (PI-C)

In other words, the agent of the giving is socially close to the speaker, while the beneficiary is socially higher than the agent and is being treated respectfully by the speaker. It is also possible to use this category when the agent is not socially close to the speaker, but evidently Japanese speakers are not completely comfortable about the choice in that case; nevertheless, there is no other category available.

One way in which VAT might be able to find answers to questions regarding social relationships is through the occurrence in the text of categorizations that entail such relationships. For example, the occurrence of an instance of UC-"SENSEI" in example 92) entails a socially higher status for the PI thus categorized than for the PIs who are this teacher's students. It thus leads to the choice of UC-"KUDASAR-". Kinship terms also provide examples of automatically entailed social status. If we take a PI that is an instance of UC-"OTOOSAN" 'father', for example, there are entailments of the following sort:

101) PI-A C→ UC-"OTOOSAN"
E→
VB-FATHER-OF (PI-A, PI-B)
VB-HIGHER-THAN (PI-A, PI-B)

That is, PI-A must be the father of someone (PI-B), and will be socially higher than that someone. It will also be the case that:

102) VB-FATHER-OF (PI-A, PI-B)
SP-SPEAKER (PI-A)
E>
VB-CLOSE-TO-SPEAKER (PI-B)

That is, if the PI-A who is the father of PI-B is at the same time the speaker, PI-B will be socially close to the speaker. The entailments derived from both 101) and 102) are relevant to the choice of a translation for English UC-"GIVE", as sketched above.

So far all our examples of translation problems have involved categorization. Certainly, however, there are also problems which arise in subconceptualization, and in the associated application of syntactic processes which lead to clause and sentence formation. We have not paid as much attention to questions of this sort, since for the most part we have been able to translate sentence for sentence with reasonable success. One example which seems fairly clear arose early in our investigation, and will be repeated here as an illustration of the challenges which are likely to arise in this respect.

At issue is the translation of the English sentence in 103), the first sentence of a fable, into the sequence of two Japanese sentences in 104):

103) There was once a wolf who saw a lamb drinking at a river, and wanted to create an excuse to eat it.

104) Mukasi aru tokoro ni kawa de mizu o nonde
once certain place in river at water drinking

iru ko-hituzi o mituketa ippiki no ookami ga imasita.
be lamb saw one wolf was

Sosite sono ookami wa sono ko-hituzi o taberu tame no
and that wolf that lamb eat for

iiwake o tukuri-ta-gatte imasita.
excuse make-want-seeming was

The question we are concerned with is why it is desirable for the Japanese translation to create two sentences where the English had only one.

We may note first of all that the English sentence contains two conjoined relative clauses ("who saw...and wanted..."). Japanese relative clauses differ syntactically from those in English in being preposed to the noun they modify. Hence, if the Japanese were to preserve the structure of the English in a single sentence, the speaker would have to say everything that the wolf saw and wanted before he ever was able to mention the wolf. The subject of the seeing and the wanting would be held in suspense for so long that addressee or reader might have some problem in interpreting what was being said. Another reason for not repeating the English structure of two relative clauses has to do with the beginning of the next sentence: in English, "For that purpose...he accused the lamb of stirring up the water..." The referent of <u>that purpose</u> in English is clear. It refers to the immediately preceding relative clause: "wanted to create an excuse to eat it." His wanting to create this excuse was his

purpose for accusing the lamb. In Japanese, however, if the clause in question were preposed to ookami (which would then be followed by the main verb of the sentence, imasita), the referent of that purpose would no longer be clear. By making the clause about the wolf's wanting to create the excuse into an independent sentence, the Japanese is able to refer to it directly at the beginning of the next sentence without difficulty.

We have not yet formalized the processes by which VAT would decide to create two sentences in the translation where the source verbalization has one. Evidently principles such as the following must eventually be built into VAT. First, there must be a restriction of some kind on the amount of material that can be included in a preposed relative clause, and perhaps especially in a relative clause that introduces the main character of a story (whose introduction cannot be put off for too long). Second, there is a need for a sentence-introductory phrase like for that purpose to have a clear referent which immediately precedes it. The task of introducing such principles into VAT's operations is formidable, but not impossible of accomplishment.

X. Future Work

It will be obvious to anyone who has tried to deal with the sorts of information and processes mentioned in this report that we have only inserted a few pin pricks into a gigantic monster whose eventual conquest calls for years of patient work. Without pretending to cover everything that needs to be done, we summarize below some of the more obvious lines of research that the report suggests.

(1) During subconceptualization we make use of statements like CJ-YIELDS (CC-1002, CC-1003). We need to extend and clarify the set of relations to which CJ-YIELDS belongs: the relations which exist between the various conceptual chunks of a text, whether these chunks be large or small.

(2) Such relations have surface consequences, of the kind illustrated in examples 8), 9), 19), and 22) above. Such consequences are in fact quite varied and subtle, being dependent in part on complex contextual considerations. Their clarification calls for extensive textual analysis.

(3) We now have a primitive device for introducing digressions and parentheses into the subconceptual hierarchy. We need to look at digressions in greater detail to determine more precisely what constitutes a digression, how best to formalize the processes by which digressions are introduced, and how they

are expressed under various conditions.

(4) During the initial period of this project we spent some time investigating the manner in which various textual genres constrain the processes of verbalization. It will be necessary to return to this question with a view toward constructing conceptual "grammars" of scientific articles, news reports, stories of various kinds, etc.

(5) Although the general nature of the framing process seems to be understood, its details need refinement. The best inventory of "case" relations has not yet been established, nor has the interaction between cases and other statuses which PIs may have, such as topic or given information.

(6) The treatment of "modifiers" (adjectives, relative clauses, adverbs) is presently oversimplified. More work on their introduction and expression is called for.

(7) The treatment of "inflections" (tense, aspect, articles, number, and the like), though it has been given a fair degree of attention already, needs to be expanded and extended.

(8) At present, if a PI has a proper name we treat it as a unique name and use it for the lexicalization of the PI whenever the latter is not "given". This procedure ignores the many interesting constraints which govern the choice among competing proper names for the same PI. Investigation of this area is

dependent on a more detailed understanding of a variety of interpersonal relationships.

(9) When a PI does not have a proper name and is not pronominalized, it must be categorized in some way that will lead to lexicalization in terms of a common noun. The factors which influence such categorization are of basic psychological interest, involving such questions as whether conceptual "features" are adequate to account for how a particular PI is categorized, and the extent to which continuous degrees of codability must be recognized. These factors will have to be included eventually in the lexicon, so that this problem is really the problem of how the lexicon should be developed.

(10) A practical problem involves the procedures by which lexical entailments are utilized. Should all the entailments associated with every item in a text be specifically created by the program, or should they somehow be held in some latent condition until they are needed? It is important to avoid the mushrooming of entailments beyond necessity, but exactly how it can be avoided is not yet clear.

(11) At present, if a PI is "given" it is automatically pronominalized. We know that pronominalization is influenced by other factors; for example, it will often not take place if ambiguity is likely to result. Such factors as a search for possible ambiguity will have to be introduced into the system.

(12) Among the "readjustments" (section VII) which are applied after a sentence has been produced, we have dealt with three types: the introduction of givenness, the introduction of the relation IDENTIFIES between a category and a PI, and the creation of a CC which represents the production of the sentence as an event in itself. Other kinds of readjustments--that is, changes in discourse information which result from the production of a sentence--need to be investigated.

(13) Our surface structure format at present consists of a series of statements representing the major components of a sentence, with all necessary surface information included. We have designed such structures so that they can be algorithmically converted into sequences of words and sentences--that is, into a normally readable text. The algorithm needs to be specifically implemented.

(14) With reference to the translation component in particular, it will be necessary to look into differences in the way different languages subconceptualize various kinds of content, differences in the treatment of various genres, differences in the placement of sentence boundaries, and so on.

(15) The construction of an extensive interlingua list (a dictionary of direct category correspondences) between pairs of languages is called for.

(16) Procedures involved in the use of entailments for the

discovery of indirect category correspondences between two languages must be refined. Aside from the need to build up a large and detailed working lexicon for each language with this goal in mind, it will be necessary to find ways of optimizing the search for corresponding entailments when no direct correspondence in interlingua is found.

(17) With respect to parsing, we need to make precise the techniques by which textual clues are utilized in the reconstruction of the verbalization processes by which the text was created. These clues are many and varied, differing to some extent from language to language, and again a large-scale empirical investigation is called for.

Attacks on these problems are appropriate on at least four fronts. First, investigators should undoubtedly continue to engage in traditional "armchair" linguistics, involving cogitation and discussions by persons steeped in the languages, procedures, and theoretical issues involved. Second, one can adapt and exploit whatever materials relevant to these questions can be found in the literature on these languages, on linguistics, on artificial intelligence, on text structure, and so on. Third, it will be possible to develop new facts through experimental work. As an example, one can investigate specific examples of human translation in order to establish ranges of variation in different verbalizations of what is essentially the same content, and to determine the optimum correspondences between

two languages in specific cases. Finally, of course, the development of a computer system can proceed in parallel with these other lines of research, handling the ever-increasing complexity in a way that the computer is uniquely suited for, and providing the indispensible testing ground for each new feature or process that is hypothesized.

Footnotes

[1]Roger W. Brown and Eric H. Lenneberg, "A Study in Language and Cognition," Journal of Abnormal and Social Psychology 49: 454-462 (1954).

[2]Wallace L. Chafe, "Language and Consciousness," Language 50:111-133 (1974).

[3]Cf. the discussion of "deep" memory in Chafe, "Language and Memory," Language 49:261-281 (1973).

[4]What follows is based on the analysis by Susumo Kuno in his The Structure of the Japanese Language (MIT Press, 1973), chapter 9.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER: RADC-TR-74-271	2 GOVT ACCESSION NO. / 3 RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle): AN APPROACH TO VERBALIZATION AND TRANSLATION BY MACHINE	5 TYPE OF REPORT & PERIOD COVERED: Final Report 1 Jun 72 - 31 May 74
	6 PERFORMING ORG. REPORT NUMBER: None
7 AUTHOR(s): Dr. Wallace L. Chafe	8 CONTRACT OR GRANT NUMBER(s): F30602-72-C-0406
9 PERFORMING ORGANIZATION NAME AND ADDRESS: The University of California at Berkeley Department of Linguistics Berkeley, California 94720	10 PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Program Element 62702F Job Order No. 45940805
11 CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (IRDT) Griffiss Air Force Base, New York 13441	12 REPORT DATE: October 1974 / 13 NUMBER OF PAGES: 114
14 MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Same	15 SECURITY CLASS. (of this report): UNCLASSIFIED / 15a DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A

16 DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17 DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18 SUPPLEMENTARY NOTES

None

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

Computational Linguistics
Machine Translation
Lexicography
Computer Programming
Artificial Intelligence

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

The report documents performance on a 24 month R&D effort oriented toward the development of a computerized model for machine translation of natural languages. The model is built around a set of procedures called verbalization, intended to simulate the processes employed by a speaker or writer in turning stored information into words. Verbalization is seen to consist of subconceptualization and lexicalization processes which involve creative choices on the part of the verbalizer, together with algorithmic

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract (Cont'd)

syntactic processes determined by the language being used. Translation is viewed as (1) the reconstruction of the verbalization processes which went into the original source language text and (2) the application of parallel verbalization processes in the target language. The target language verbalization looks for creative choices to the source language verbalization and tries to apply corresponding choices simultaneously with application of syntactic processes dictated by the grammar of the target language. Verbalization and translation processes are illustrated in some detail with examples taken from English and Japanese. Some of these processes have been implemented in an interactive program on CDC 6600 at the Lawrence Berkeley Laboratory (AEC), but the main intent of the report is to demonstrate the kinds of processes that need to be incorporated in such a system.

MISSION
of
Rome Air Development Center

RADC is the principal AFSC organization charged with planning and executing the USAF exploratory and advanced development programs for electromagnetic intelligence techniques, reliability and compatibility techniques for electronic systems, electromagnetic transmission and reception, ground based surveillance, ground communications, information displays and information processing. This Center provides technical or management assistance in support of studies, analyses, development planning activities, acquisition, test, evaluation, modification, and operation of aerospace systems and related equipment.

Source AFSCR 23-50, 11 May 70